Susanne Milcher and Manfred M. Fischer

On labour market discrimination against Roma in South East Europe

Paper

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 directs interest on country-specific labour market discrimination Roma may suffer in South East Europe. The study lies in the tradition of statistical Blinder-Oaxaca decomposition analysis. We use microdata from UNDP's 2004 survey of Roma minorities, and apply a Bayesian approach, proposed by Keith and LeSage (2004), for the decomposition analysis of wage differentials. This approach is based on a robust Bayesian heteroscedastic linear regression model in conjunction with Markov Chain Monte Carlo (MCMC) estimation. The results obtained indicate the presence of labour market discrimination in Albania and Kosovo, but point to its absence in Bulgaria, Croatia, and Serbia.

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

JEL classification: J15, J31, J71, C11, O52

***Corresponding author**
Professor Manfred M. Fischer, Manfred.Fischer@wu.ac.at
Institute for Economic Geography and GIScience,
Vienna University of Economics and Business
Nordbergstr. 15/4/A, A-1090 Vienna, Austria
phone: +43-1-31336-4836 fax: +43-1-31336-703

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 of Roma in South East Europe have settled, 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.

The challenges for the Roma minority are well known: overcoming poverty, increasing access to education, and diminishing labour market discrimination. But despite a general awareness of labour market discrimination of Roma in these countries, information on labour market discrimination needed for policy actions is scarce, fragmented and often anecdotal. This is due to several reasons. First and foremost, the simple question, “who is Roma?”, does not have a simple answer, given the different meanings ascribed to the notion of Roma, and the diversity of the Roma universe. Ethnographers have, for example, identified 60 different groups in

Bulgaria (Revenga et al. 2002), and such diversity may also exist in other countries. In addition to these ethnical differences, there is significant diversity among Roma settlements: rural versus urban, integrated versus non-integrated, homogenous versus heterogeneous, and affiliations with different religious denominations (Muslims versus Christians). Some groups speak variations of the Roma language while others do not (Revenga et al. 2002). As a result, it is difficult to identify Roma based upon distinctive characteristics, such as appearance, language or family names[1].

This study uses survey data collected from face-to-face interviews with 9,889 Roma respondents in Albania, Bulgaria, Croatia, Kosovo and Serbia. These data come from a data collection exercise performed by UNDP's survey of Roma minorities and other vulnerable groups, conducted in October 2004. This survey took a multifaceted approach to the issue of ethnicity, including questions on self-identification, interviewer identification, language, and parent's language. In a number of respects this survey is unique in its scale and consistency over the five countries considered in this study. The data for each country are comparable because they are based on a common questionnaire (translated into respective local languages) and on identical sampling design methodology.

The focus of this study is on a specific form of labour market discrimination, known as wage discrimination that exists when the relative wages of non-Roma exceed the relative wages that would have prevailed if Roma and non-Roma were paid according to the same criteria[2]. This form of discrimination may be studied in terms of statistical decomposition analysis. Since its

---

[1] Note that one's self-identification with a certain ethnic minority, such as Roma is not equal to her/his perceived belonging to such minorities. Perceived ethnic origin and self-identity are rather different notions.

[2] Other forms of labour market discrimination, for example, stemming from occupational barriers, are out of the scope of this study.

popularization by Blinder (1973) and Oaxaca (1973), wage decomposition methodology has become the standard approach to estimating the extent of labour market discrimination on the basis of gender, race, and ethnicity (see, for example, Patrinos and Sakellariou 1992, Kimmel 1997, Oaxaca and Ransom 1994, MacIsaac and Patrinos 1995, Maani 2002). Decomposition analysis explains wage differentials in terms of differences in individual characteristics (characteristics effect) and differences in the ordinary least-squares coefficients of wage regression estimates (coefficients or discrimination effect).

Our study departs from standard Blinder-Oaxaca decomposition analysis of wage differentials by using a Bayesian approach to statistical inference for both discrimination and characteristics effects estimates. This approach suggested by Keith and LeSage (2004), is based on Markov Chain Monte Carlo (MCMC) estimation of a robust Bayesian heteroscedastic linear wage regression model, and shows several advantages over the traditional least-squares method of wage decomposition. *First*, MCMC estimation provides a simple and easy method for obtaining the posterior distributions of the characteristics and discrimination effects needed for testing their significance. Obtaining these posterior distributions without relying on Bayesian MCMC estimates is a difficult task since the closed forms of these distributions are not well defined (Radchenko and Yun 2003). *Second*, 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). *Finally*, degradation in precision of the characteristics and discrimination effects that typically accompany least-squares estimates in the presence of heteroscedasticity or outliers can be avoided (Keith and LeSage 2004).

The remainder of the paper is organised as follows. Section 2 briefly reviews the standard approach to the Blinder-Oaxaca decomposition of wage differentials. In Section 3 we discuss the Bayesian approach along with some details on the MCMC estimation methodology. Section 4 describes the survey data and variables used for the analysis, and Section 5 presents the paper's empirical findings. Finally, Section 6 offers some closing remarks.

2 The standard Blinder-Oaxaca decomposition of wage differentials

In what follows, we refer to the group of Roma that suffers labour market discrimination as j=1, and the group of non-Roma that suffers no discrimination[3] as j=2. Decomposition analysis assumes that if there were no discrimination, the wage structure currently faced by non-Roma would also apply to Roma. This assumption says that non-Roma would on average receive in the absence of discrimination the same wages as they presently receive, but that discrimination takes the form of Roma receiving less than a non-discriminatory labour market would award them.

Ordinary least-squares estimation of a wage equation for any given group j of workers provides an estimate of the wage structure applicable to that group j. The wage equation to be estimated separately for each group j has the semi-log functional form given by

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \qquad j \in \{1,2\} \tag{1}$$

[3] The non-Roma (majority) group is used as reference.

where $\mathbf{Y}_j$ denotes the *nj*-by-1 vector of log-wages for *nj* workers in group *j*. The matrix $\mathbf{X}_j$ contains $k-1$ column vectors representing worker characteristics (such as job experience and education) that purport to explain wage variation over the two samples of Roma and non-Roma workers, as well as a column vector of ones related to the intercept. The *k*-by-1 parameter vector $\boldsymbol{\beta}_j$ provides a measure of the responsiveness of wages to the various characteristics for the two demographic groups, and a constant. The disturbance vector $\boldsymbol{\varepsilon}_j$ is typically assumed to follow a zero mean, constant variance normal distribution.

If Eq. (1) is estimated separately for cross-section samples of Roma (*j*=1) and non-Roma (*j*=2), then since regression lines pass through the means of the variables we get

$$\overline{Y}_2 - \overline{Y}_1 = \left(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1\right)\hat{\boldsymbol{\beta}}_2 + \overline{\mathbf{X}}_1\left(\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1\right) \tag{2}$$

where $\overline{Y}_1$ and $\overline{Y}_2$ denote the sample means of the vectors $\mathbf{Y}_j$ (*j*=1, 2), and $\overline{\mathbf{X}}_j$ (*j*=1, 2) are 1-by-*k* vectors containing the means of the *k* variables for Roma and non-Roma workers, and $\hat{\boldsymbol{\beta}}_j$ (*j*=1, 2) are the consistent estimates of $\boldsymbol{\beta}_j$ estimated by OLS.

The first term on the right-hand side of Eq. (2) is the part of the log-wage differential due to different (average) characteristics of Roma and non-Roma, $C = (\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)\hat{\boldsymbol{\beta}}_2$, known as the "characteristics effect". The second term, $D = \overline{\mathbf{X}}_1(\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1)$ is the part of the differential due to different coefficients, or different wage structures. 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 merely to differences in pay-related characteristics then this second term can be interpreted as the part of the log-wages differential due to discrimination. This is the essence of the Blinder-Oaxaca decomposition approach.

Interest in this paper focuses on inferences regarding the coefficients effects term *D*, which provides a revealed preference view of the way in which Roma workers characteristics are valued (by employers) relative to non-Roma workers during wage determination. The characteristics effects term *C* may be viewed simply as a control variable and may be useful for inferences concerning which characteristics exert a significant impact on wage determination.

While this decomposition holds a great deal of intuitive appeal, it is less clear how one should go about drawing an inference regarding the statistical significance of these effects. These effects have extremely complicated statistical distributions reflecting the manipulation used to produce the decomposition. Moreover, inferences are likely to be sensitive to maintained regression hypotheses such as the assumption of homoscedastic disturbances, a lack of omitted variables and simultaneity bias, etc. (LeSage and Charles 2008).

Oaxaca and Ransom (1998) provide 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 for testing the significance of the two effects. These assumptions may not be valid in the face of outliers and small samples likely to be characterized by heteroscedasticity.

3 The Bayesian approach

As with the standard Blinder-Oaxaca decomposition analysis based on OLS estimates, a Bayesian approach also separately estimates the regression wage equation for groups j=1 (Roma) and j=2 (non-Roma). Bayesian MCMC estimation can be applied to generate a large sample of MCMC draws for the parameter vectors $\boldsymbol{\beta}_j$ (j=1, 2) that reflect the entire posterior distribution for these parameters. These draws can be used to construct the complete posterior distributions for the characteristics and discrimination effects that are of interest in the wage differential decomposition.

We follow Keith and LeSage (2004) to use a robust Bayesian heteroscedastic variant of the basic linear wage regression model given by Eq. (1), in conjunction with MCMC estimation, for the decomposition analysis. This Bayesian variant of the regression model given by

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \qquad j = \{1,2\} \tag{3a}$$

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}(0,\, \sigma_j^2 \mathbf{V}_j) \qquad j = \{1,2\} \tag{3b}$$

$$\mathbf{V}_j = \operatorname{diag}(v_1,\ldots,v_{nj}) \qquad j = \{1,2\} \tag{3c}$$

introduces a set of variance scalars $(v_1,\ldots,v_{nj})$ for each of the two wage equations, representing unknown parameters to be estimated. The generalization of the conventional assumption of normal constant variance disturbances allows controlling for outliers and heteroscedastic variances across samples of n_1 and n_2 workers.

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

$$\pi\,(\boldsymbol{\beta}_j) \sim \mathcal{N}(\mathbf{c}_j, \mathbf{T}_j) \tag{4a}$$

$$\pi\,(r / v_j) \sim \text{IID } \chi^2(r) \tag{4b}$$

$$\pi(1/\sigma_j^2) \sim \Gamma(d_j, v_j)\,. \tag{4c}$$

$$r \sim \Gamma(m, h) \tag{4d}$$

Given our interest in drawing inferences regarding $\boldsymbol{\beta}_j$ based on the sample data, non-informative rather than informative prior assignments seem to be reasonable for the parameters $\boldsymbol{\beta}_j$ and σ_j. $\boldsymbol{\beta}_j$ is assigned a *normal* conjugate prior, which can be made almost diffuse by setting the vector of the prior means $\mathbf{c}_j = 0$ and the prior variance-covariance $\mathbf{T}_j = \mathbf{I}_k \cdot 1\mathrm{e}+10$ where e denotes the mathematical constant e (Euler's number). 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 is associated with setting the parameters $d_j = v_j = 0$ in Eq. (4c).

Prior information concerning the variance scalars v_j that arise in the two wage 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 $\mathbf{V}_j$ by adding only a single parameter (*r*) to the model. Note that we will use the same value for this hyperparameter for both wage 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. Using m=8 and h=2 would assign a prior to r centred on a small r=4 with variance of r equal to two. 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 $\boldsymbol{\beta}_j, \sigma_j$ 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 $\boldsymbol{\beta}_j$ takes the form of a multivariate normal with mean and variance-covariance given by

$$\boldsymbol{\beta}_j \mid \left(\sigma_j, \mathbf{V}_j\right) \sim \mathcal{N}\left\{\mathbf{H}_j\left(\mathbf{X}_j' \, \mathbf{V}_j^{-1} \, \mathbf{Y}_j + \sigma_j^2 \, \mathbf{T}_j^{-1} \mathbf{c}_j\right), \sigma_j^2 \mathbf{H}_j\right\} \tag{5a}$$

$$\mathbf{H}_j = \left(\mathbf{X}_j' \, \mathbf{V}_j^{-1} \, \mathbf{X}_j + \mathbf{T}_j^{-1}\right)^{-1}. \tag{5b}$$

Let $\mathbf{e}_j = \mathbf{Y}_j - \mathbf{X}_j' \boldsymbol{\beta}_j$, then the conditional posterior density for σ_j takes the form of a $\chi^2(nj)$ distribution

$$\left\{\sum_{i=1}^{nj}(e_{ji}^2 / v_{ji}) / \sigma_j^2\right\} \Big| (\boldsymbol{\beta}_j, \mathbf{V}_j) \sim \chi^2(n_j). \tag{6}$$

The posterior distribution of $\mathbf{V}_j$ conditional on $(\boldsymbol{\beta}_j, \sigma_j)$ is proportional to a $\chi^2\,(r+1)$ distribution

$$\left\{(\sigma_j^{-2}\, e_j^2 + r) / v_j\right\} \Big| (\boldsymbol{\beta}_j, \sigma_j) \sim \chi^2\,(r+1). \tag{7}$$

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. (5) through (7), we can formulate an MCMC sampler for the model by the following steps:

(i) Begin with arbitrary values for the parameters which we denote $\boldsymbol{\beta}_j^0, \sigma_j^0, v_j^0$ 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 $\boldsymbol{\beta}_j$ using Eq. (5) conditional on the initial values σ_j^0, v_j^0 and r^0.

(iii) Use the computed mean and variance of $\boldsymbol{\beta}_j$ to draw a multivariate normal random vector, labelled $\boldsymbol{\beta}_j^1$.

(iv) Compute expression (6) using $\boldsymbol{\beta}_j^1$ determined in *Step* (iii) and take this value along with a random $\chi^2(n_j)$ draw to determine σ_j^1.

(v) Using $\boldsymbol{\beta}_j^1$ and σ_j^1, compute expression (7) 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 $(\boldsymbol{\beta}_j^q, \sigma_j^q, v_j^q, r^q)$ 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. (3) produces a set of estimates that converge in the limit to the true (joint) posterior distribution of the parameters.

In addition to parameters, we are interested in the posterior distribution of the characteristics effect which can be constructed using draws q=1, ... as $(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)\,\overline{\boldsymbol{\beta}}_2^q$ and the coefficients effects found using $\overline{\mathbf{X}}_1\,(\overline{\boldsymbol{\beta}}_2^q - \overline{\boldsymbol{\beta}}_1^q)$. Statistical significance of these effects can be tested using Bayesian *p*-level calculations that are Bayesian equivalents to *t*-statistics. These calculations are based on an enumeration of the draws larger or smaller than zero, depending on the sign of the coefficient by counting the number of draws larger or smaller than zero, depending on the sign of the coefficient (see Gelman et al. 1995). One can also construct posterior credible intervals using 90 or 95 percent levels from the MCMC draws.

4 Data and variables

We use survey data collected from face-to-face interviews with 9,889 Roma and 7,438 non-Roma respondents in Albania, Bulgaria, Croatia, Kosovo and Serbia. The data come from a cross-country survey of Roma minorities and other vulnerable groups, conducted by UNDP, the UN's global development programme, in October 2004. The survey questionnaire that was used to generate the data follows the philosophy of integrated household surveys, with separate components containing both individual and household modules. Within the individual module, each household member's profile was registered (demographic characteristics, economic status, education, health). The household module addresses issues related to the household in general. Questions related to incomes and expenditures were addressed in both modules, making it possible to cross-check the interview results.

It is important to note that random sampling was not feasible due to the complexities associated with defining Roma populations. Hence, a "pyramid" sampling model was used instead. This model is based on the assumption that national census data provide reasonably adequate representations of the structure and territorial distribution of the individuals who identify themselves as Roma. Based on this assumption, the universe of Roma population was defined as Roma living in 'Roma settlements as areas of compact Roma population'. Those settlements or areas were defined as settlements where the share of Roma population equals or is higher than the national share of Roma population in the given country as reflected in the census data (see UNDP 2006 for more details).

Sampling clusters were determined taking Roma organizations' estimates of Roma population, the distribution of the settlements and population sizes into account. Respondents were identified then using a 'random route' selection process, reflecting the demographic structure of the Roma population in the respective country. The major drawback of this sampling methodology relates to the neglect of Roma living in municipalities where the share of Roma in the total population is below national averages. Thus, the samples are not fully representative for the entire Roma populations of the countries covered in the survey. But the data generated by these samples are broadly consistent with census data, since this survey's data are based on relative numbers (economic and demographic structure, and regional distribution) instead of absolute numbers of Roma registered in the censuses. In order to derive data for meaningful comparisons, control groups' samples of non-Roma populations were constructed in each country using similar procedures as for the Roma samples[4].

We used the 9,889 Roma and 7,438 non-Roma respondents in the five countries to generate a sample of hypothetically logged wage incomes for 841 Roma and 1,792 non-Roma workers. These samples of 16-65 years aged individuals were obtained by excluding (i) self-employed and others not working, (ii) employed in the agricultural sector, (iii) those with missing wage income due to not working or working without pay (for example in subsistence agriculture), (iv) those working in the shadow economy (begging, gambling) or receiving state benefits as primary source of income, and (v) those with missing data on some subset of independent variables.

[4] The Roma and non-Roma samples were generally drawn from the same municipalities or administrative units. In some municipalities with a very high share of Roma population, however, the share of non-Roma population was not sufficiently large for using 'random route' selection processes. In such cases (such as isolated Roma settlements or segregated neighbourhoods), the non-Roma sample was based on a typologically similar settlement in the same district (administrative unit) with a Roma population equal to or higher than the national average. The criterion for choosing this settlement was that it be the 'closest village accessible by road connection' (UNDP 2006).

The country-specific sample sizes are small, especially those for Roma workers[5]. But they are not too small for decomposition analysis. The differences between Roma and non-Roma sample sizes are due to smaller proportions of Roma with wage income as a consequence of much higher unemployment rates. In studies on ethnic wage discrimination smaller sample sizes appear to be the rule rather than the exception (see, for example, Patrinos and Sakellariou 1992, MacIsaac and Patrinos 1995).

Position Table 1 about here

The UNDP survey does not provide information on actual wages but on income. Income may include diverse sources of non-labour income. But the construction of the Roma and non-Roma samples described above justifies using income as a proxy for wage income. The choice of the independent variables is limited by the constraints of data availability. We use six independent variables to specify the matrix $\mathbf{X}_j$ (j=1, 2) in Eq. (3a). 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. Since data on the actual number of years of work experience are not available, we use age as reasonable proxy for potential work experience. In accordance with the post-schooling investment model of human capital formation as developed in Mincer (1974), a quadratic experience variable is in the wage equations. The corresponding

[5] Albania: 289 Roma and 570 non-Roma; Bulgaria: 241 Roma and 370 non-Roma; Croatia: 77 Roma and 219 non-Roma; Kosovo: 123 Roma and 280 non-Roma; Serbia 111 Roma and 353 non-Roma.

coefficient measures the combined effects of the average rate of return to on-the-job training and the length of the investment horizon. 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 male dummy is taken to control for gender-specific effects[6].

Table 2 presents descriptive statistics of the dependent and independent variables. In all the countries the mean log wage income for Roma is lower than that for non-Roma. With 0.70 the differential is largest in Albania and with 0.26 lowest in Croatia. The mean values for education, work experience and the squared experience variable are consistently higher for non-Roma. Greater proportions of Roma workers appear to be engaged in low skilled, low quality forms of employment.

Position Table 2 about here

5 Empirical results

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[7]. The first four columns

---

[6] If interest is focused on gender discrimination among Roma, it would be necessary to separately estimate wage equations for Roma women and Roma men. The small Roma sample sizes do not allow pursuing this interesting question further in the context of this study.

[7] Public domain algorithms in the MATLAB matrix programming language that implement the estimation methodology can be found in LeSage's Econometrics Toolbox at www.spatial-economtrics.com.

present the parameter estimates of the Bayesian semi-log wage 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[8]. 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.

Space limitations allow discussion of only a few selected aspects of the regression coefficients reported. 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 wage income in all countries, but the impact is generally relatively low and not significant in Serbia. Working full time and in a skilled occupation appears to be most important for increasing the wage income 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 wage income differentials (in log terms) into characteristics and coefficients (discrimination) effects, based on the Bayesian MCMC estimation methodology. The Bayesian estimates

[8] It is worth noting that the decomposition based on OLS estimates shows similar results (see appendix), which is not surprising considering that we use a normal-diffuse prior for the *beta* parameters. The standard deviations (computed using the Oaxaca and Ransom 1998 asymptotic variance calculation) point to a larger dispersion for the least squares estimates than those of the robust heteroscedastic Bayesian regression model. Degraded precision in the estimates exerts an adverse impact on the asymptotic normal approximation to the variance of the discrimination effects 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 effects, $\mathrm{H0}_C: C = 0$, and no discrimination effects, $\mathrm{H0}_D: D = 0$.

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 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 wage 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 wage income differential. The lower level of education and lower share of Roma in full time work compared to non-Roma 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 wage income gap in Albania.

- *Bulgaria*: About 90 percent of the log wage 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 the lower endowments of Roma do indeed explain a large fraction of the observed differences in wage income between non-Roma and Roma groups in this country. Much of this reflects huge differences in educational endowments and access to education, most likely caused by lower quality schooling of Roma in general and segregated schooling in particular.

- *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 wage 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 the quadratic experience variable that captures decreasing marginal returns to work experience. Note that these discrimination effects appear to matter most, suggesting that Roma experience different returns to work experience than non-Roma do and this contributes to a widening of the wage income gap.

- *Kosovo*: The aggregate characteristics and coefficients effects explain 32.2 and 67.8 percent of the log wage 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 male) and one individual discrimination effect (full time) are statistically significantly different from zero. The full time variable largely contributes to widening the wage 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 wage income gap between Roma and non-Roma is largely explained through the lower levels of education of Roma and differing returns to education for Roma. Thus, bringing the education level of Roma in Serbia to the level of non-Roma would substantially reduce the wage income differential.

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 male) 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).

6 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 wage income differentials among Roma and non-Roma population groups in five South East European countries, using sub-samples from the 2004 UNDP survey.

This approach has several merits. One is that the posterior distributions of the characteristics and discrimination effects are easily obtained by using Markov Chain Monte Carlo sampling. 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. Variance estimates derived from MCMC estimation are known to reflect the true posterior variance given a sufficiently large sample of MCMC draws. Last but not least, degradation in precision of the discrimination effects that typically accompany least-squares estimates in the presence of heteroscedasticity can be avoided.

The results obtained suggest the presence of statistically significant discrimination effects in Albania and Kosovo, but their absence in Bulgaria, Croatia and Serbia. The discrimination effects explain 67.8 and 42.5 percent of the wage income differential between Roma and non-Roma in wage employment in Kosovo and Albania, respectively. Labour market discrimination is apparently an important factor in explaining wage income differences among Roma and non-Roma groups that are in paid market work in these two countries. But differences in measured characteristics (especially education) and not wage discrimination against Roma appear to be important reasons for the shortfall in incomes for Roma in wage employment 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 (1995) *Bayesian data analysis*. Chapman and Hall, London

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

Kimmel J (1997) Rural wages and returns to education: Differences between whites, blacks, and American Indians. *Economics of Education Review* 16(1): 81-96

LeSage JP, Charles JS (2008) Using home buyers' revealed preferences to define the urban-rural fringe. *Journal of Geographical Systems* 10(1): 1-21

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

Maani SA (2002) Education and Maori relative income levels over time: the mediating effect of occupation, industry, hours of work and locality. *Working Paper no. 02/17*. New Zealand Treasury, New Zealand

MacIsaac D, Patrinos H (1995) Labor market discrimination against indigenous people in Peru. *Journal of Development Studies* 32(2): 218-33

Mincer J (1974) *Schooling, experience, and earnings.* National Bureau of Economic Research and Columbia University Press, New York

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

Patrinos HA, Sakellariou CN (1992) North American Indians in the Canadian labour market: A decomposition of wage differentials. *Economics of Education Review* 11(3): 257-266

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

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

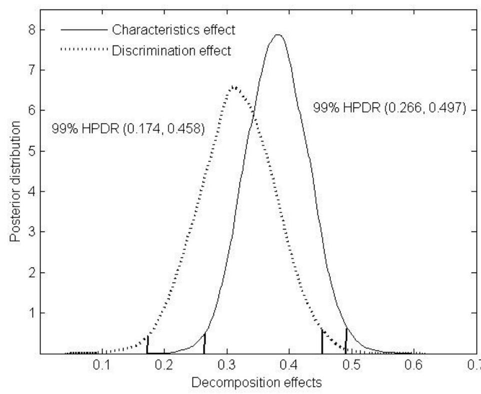
UNDP (2006) *At risk: Roma and the displaced in Southeast Europe*. United Nations Development Programme, Regional Bureau for Europe and the Commonwealth of Independent States, Bratislava

List of figures

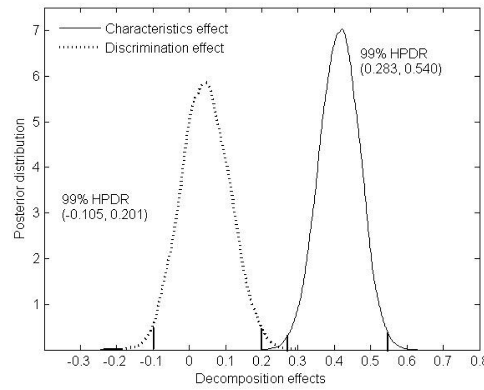
List of tables

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

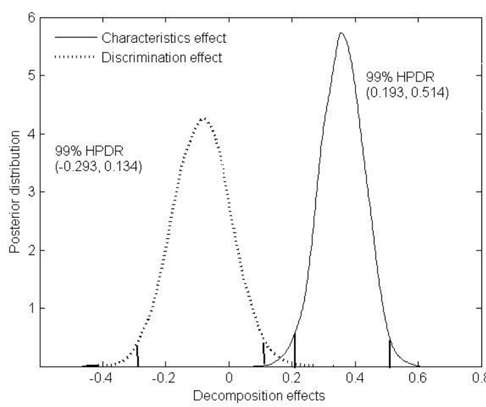
(a) Albania



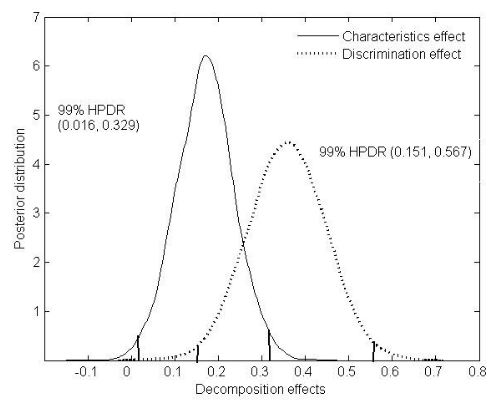
(b) Bulgaria



(c) Croatia



(d) Kosovo



(e) Serbia

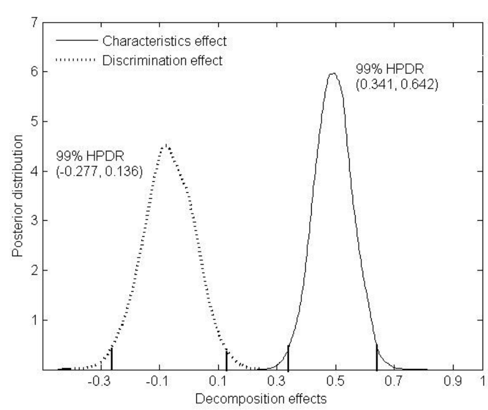


Table 1 Variables used in the analysis

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 [potential work experience]
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
Male	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
Male										
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. squared	-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)	
Male	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. squared	-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)	
Male	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 (n1=77, n2=219)								
Constant	3.329 (0.000)	0.628	4.225 (0.000)	0.430			0.896 (0.246)	0.768
Education	0.052 (0.005)	0.020	0.074 (0.000)	0.012	0.318 (0.000)	0.051	0.188 (0.352)	0.201
Work exp.	0.100 (0.003)	0.035	0.023 (0.121)	0.020	0.121 (0.248)	0.104	-2.493 (0.061)	1.320
Work exp. squared	-0.130 (0.007)	0.052	-0.018 (0.231)	0.024	-0.074 (0.459)	0.100	1.283 (0.056)	0.666
Full time	0.089 (0.338)	0.200	0.380 (0.014)	0.166	0.023 (0.024)	0.010	0.253 (0.260)	0.224
High skills	0.243 (0.020)	0.120	-0.028 (0.418)	0.127	-0.013 (0.829)	0.062	-0.119 (0.124)	0.077
Male	0.215 (0.030)	0.115	0.085 (0.069)	0.057	-0.015 (0.140)	0.010	-0.093 (0.312)	0.092
Aggregate					0.359 (0.000)	0.069	-0.085 (0.360)	0.092
(d) Kosovo (n1=123, n2=280)								
Constant	2.971 (0.000)	0.623	3.589 (0.000)	0.404			0.618 (0.405)	0.740
Education	0.022 (0.100)	0.017	0.012 (0.189)	0.014	0.059 (0.379)	0.066	-0.072 (0.642)	0.155
Work exp.	0.067 (0.021)	0.033	0.059 (0.002)	0.021	0.165 (0.005)	0.057	-0.284 (0.837)	1.375
Work exp. squared	-0.093 (0.016)	0.043	-0.073 (0.002)	0.025	-0.159 (0.004)	0.055	0.262 (0.701)	0.680
Full time	0.658 (0.000)	0.105	0.125 (0.085)	0.092	0.035 (0.176)	0.026	0.286 (0.000)	0.075
High skills	0.436 (0.000)	0.111	0.222 (0.002)	0.077	0.091 (0.004)	0.032	-0.057 (0.117)	0.036
Male	0.088 (0.305)	0.173	0.284 (0.001)	0.091	-0.020 (0.002)	0.006	0.177 (0.316)	0.176
Aggregate					0.170 (0.012)	0.067	0.358 (0.000)	0.089

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 (n1=111, n2=353)								
Constant	3.344	0.662	3.288	0.404			-0.056	0.776
	(0.000)		(0.000)				(0.943)	
Education	0.016	0.018	0.080	0.010	0.314	0.040	0.593	0.192
	(0.187)		(0.000)		(0.000)		(0.002)	
Work exp.	0.029	0.035	0.001	0.018	0.001	0.025	-1.085	1.532
	(0.200)		(0.478)		(0.954)		(0.479)	
Work exp. squared	-0.016	0.045	0.005	0.022	0.005	0.025	0.345	0.827
	(0.359)		(0.415)		(0.833)		(0.677)	
Full time	0.389	0.119	0.561	0.123	0.147	0.032	0.116	0.114
	(0.001)		(0.000)		(0.000)		(0.311)	
High skills	0.323	0.120	0.161	0.098	0.075	0.046	-0.076	0.072
	(0.005)		(0.051)		(0.102)		(0.294)	
Male	0.072	0.139	0.188	0.050	-0.051	0.014	0.095	0.122
	(0.297)		(0.000)		(0.000)		(0.436)	
Aggregate					0.492	0.066	-0.068	0.089
					(0.000)		(0.441)	

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

Appendix: Decomposition analysis based on OLS and Bayesian MCMC estimates: Aggregate discrimination effects

	OLS-Oaxaca-Ransom		Bayesian	
	Discrimination effect (*p*-value)	Std. Err.	Discrimination effect (*p*-value)	Std. Err.
Albania	0.233	0.075	0.317	0.061
	(0.002)		(0.000)	
Bulgaria	0.104	0.072	0.045	0.067
	(0.148)		(0.502)	
Croatia	-0.044	0.106	-0.085	0.092
	(0.676)		(0.360)	
Kosovo	0.496	0.128	0.358	0.089
	(0.000)		(0.000)	
Serbia	-0.064	0.107	-0.068	0.089
	(0.550)		(0.441)	