Authors: Gréta Zsikla & Márton Nagy
Data Analysis 2 & Coding 1 joint 2nd assignment

Labour opportunities of Romani people in Hungary: the case of communal workers

1. Introduction

The disadvantage of Romani people on the Hungarian labour market is a well-known and much studied phenomenon. However, it remains an open question whether this is a result of discrimination or the over-representation of Romani people in marginalized social groups. Our paper seeks to examine this problem in the context of communal work (*közfoglalkoztatás* in Hungarian). We study the question whether there is a relationship between the number of Romani people per capita in each Hungarian municipality in 2022[1] and the number of communal workers per employed people, controlling for a number of municipality-level socio-economic and development indicators. Our hypothesis is that there is an initial positive association when no controls are included, but we expect this relationship to disappear as controls are introduced. The acceptance of the hypothesis would be in line with the *over-representation* scenario, while rejecting it would point towards *discrimination,* though our results are bound to be inconclusive[2].

2. Source data

The variables in our dataset come from two sources. We have manually downloaded municipality-level indicators from TEIR[3], and augmented this with information on municipalities' legal status, county and district from a table downloaded from KSH[4]. This augmented dataset contains 33 variables in total, a detailed description of which can be found in Appx. A1.1. Our dataset provides full coverage of Hungarian municipalities in 2022, and it can also be argued that it captures a random realization of the general pattern between communal workers and the Romani population.

2.1. Data manipulation and cleaning

The raw indicators in our dataset contained a great number of missing values. For most of the variables[5], we imputed these with zeros, as (1) the minima of these variables were usually non-zero and (2) on the online data viewer of TEIR, these missing values showed up as zero. For variables related to the ratios of elementary students, we noticed that these were only missing if the number of elementary students were zero. Thus, these were *true* missing values, as they related to cases when one would have to divide by zero. So, we included a flag for the imputed variables in these cases. Also, we have dropped the capital city from the analysis, as we did not have district level data on it. This way, we ended up with 3,153 observations (out of the initial 3,155).

Next, we have created some new variables: (1) we scaled percentage variables to a 0-1 ratio scale; (2) we added ratio variables by dividing each of our variables with a sensible reference group (e.g. per capita, per employed people); and (3) we also created some dummy variables for municipalities' legal status (see details in Appx. A1.2.). Having these, we divided our sample randomly into an 80 percent train (2522 obs.) and a 20 percent test (631 obs.) sample (to perform robustness checks in Sec. 4.1.). All analysis below relates to the train sample only.

Another possible problem with our source data is that there is a chance that the Romani population is systematically underreported due to the fact that the census surveys are self-administered and Romani ethnicity is highly stigmatized in Hungary, thus some Romani individuals might not want to report themselves as such. We will get back to this issue Sec. 4.2.

---

[1]We have chosen to analyze data from 2022, as many of the needed control variables are only available in years when a general census (*népszámlálás* in Hungarian) has been conducted.
[2]See Sec 5. on the possibility of an ecological fallacy in detail.
[3]National Spatial Development and Planning Information System (*Országos Területfejlesztési és Területrendezési Információs Rendszer* in Hungarian).
[4]Hungarian Central Statistical Office (*Központi Statisztikai Hivatal* in Hungarian).
[5]We did not impute the only missing value for population aged above 7, as it was implausible that there is a municipality where only children live - we rather dropped this observation. We also did not impute variables related to the development of roads, as it was inconsistent whether the total length of roads and the length of developed and undeveloped roads was missing. Because of this, we excluded these variables from the analysis.

2.2. Descriptive and exploratory analysis

Our two main variables of interest are communal workers per employed people (outcome) and Romani population per capita (explanatory). The main issue with these variables was their right-skewed distribution with numerous zero values. For the Romani population, we decided to take logs by assigning a positive, but minimal non-zero value to observations with zero Romani population, and we added a flag to indicate whether the initial value was zero. For communal workers, adding a flag and taking the log this way was not an option, as this is a left-hand side variable. We rather opted for a two-step analysis: (1) we estimate probability models on a dummy variable denoting whether the municipality has any communal workers (so all observations can be included), and then (2) we estimate regular OLS models for the log of communal workers (but now only the observations with at least one communal worker are included). We then estimated lowess regressions for both set-ups to get an idea for the functional forms. For the histograms and lowess estimates for our main variables, see Appx. A1.3 and A1.4. We have also taken logs of some control variables which had positive minima and right-skewed distributions. All other controls were left at level. The summary statistics of all variables can be found in Appx. A1.3.

3. Regression analysis

3.1. Modeling options

The key choices we had to make during our analysis were: (1) whether to estimate LPM or logit models[6] for the binary set-up; (2) whether to include controls in their raw form, or to create principal components out of them (as some were strongly correlated[7]); and (3) whether it is worth including controls with appropriate functional forms instead of a simple linear way. For the first choice, we opted for logit models, as our sample was quite unbalanced (that is, the majority of the municipalities had communal workers), so LPM predictions were out of the 0-1 range for most of the observations. For the second, we experimented with both model types and principal component ones offered only a marginally worse fit than regular ones, so we have decided to use them to avoid collinearity issues. For the third question, lowess estimates showed that most of the controls follow a non-linear pattern, so adding functional forms improved our models' fit to some extent. These choices are supported by goodness of fit measures and plots, presented in Appx. A2.3 and A2.5.

3.2. Model choices for probability and regular set-ups

For both the binary and a regular OLS set-ups, we estimated 10 model specifications in total (see in Appx. A2.2. and A2.4.). Here, we only present our chosen models, that is the ones using principal components as controls, with functional forms added and including significant controls only - together with the simple, uncontrolled models for comparison.

Table 1: Logit models' key marginal differences (N=2,522)

	Model without controls	Model with controls
ln[Romani p.c.]	0.055***	0.026**
spline < -4.5	(0.012)	(0.011)
ln[Romani p.c.]	0.029**	-0.007
spline > -4.5	(0.013)	(0.011)
ln[Romani p.c.]	-0.007*	-0.001
spline < -4.5 x flag[8]	(0.004)	(0.004)
Brier-score	0.062	0.053

*Note: Heteroskedasticity robust standard errors (HC1) in parentheses. *p<0.1; **p<0.05; ***p<0.01*

Table 2: OLS models' key coefficients (N=2,338)

	Model without controls	Model with controls
ln[Romani p.c.]	-0.369	-0.441**
spline < -6	(0.312)	(0.212)
ln[Romani p.c.]	0.802***	0.196***
spline > -6	(0.020)	(0.017)
ln[Romani p.c.]	-0.163**	0.050
spline < -6 x flag	(0.067)	(0.046)

[6]We did not consider probit models due to their high similarity to logit results.

[7]For the collinearity analysis, refer to Appx. A2.1.

[8]You might be wondering why the flag is only included in the model as an interaction. A detailed explanation of this can be found in the introductory part of Appx. A2.

	Model without controls	Model with controls
R-squared	0.396	0.817

*Note: Heteroskedasticity robust standard errors (HC1) in parentheses. *p<0.1; **p<0.05; ***p<0.01*

3.3. Model interpretations[9]

For the logit model, the lower spline indicates that for observations having log Romani population p.c. less than -4.5 and where the initial value was not zero, observations with a ten percent higher Romani ratio have on average a 0.26 percentage point higher chance to have communal workers, all other controls being equal. This association is significant at 5%. The second spline tells us that for observations having log Romani population p.c. larger than -4.5, observations with a ten percent higher Romani ratio have on average a 0.07 percentage point lower chance to have communal workers, all other controls being equal. This association is not significant. As a 1% higher Romani population p.c. is meaningless if the actual baseline is zero, the last marginal difference is meaningless on its own. What we can look at instead is the average of predicted values when the flag is one, meaning the initial value was zero. This tells us that the average predicted probability of having a communal worker is 80.8% when the flag is 1.

For the OLS model, the lower spline indicates that for observations having log Romani population p.c. less than -6 and where the initial value was not zero, observations with a ten percent higher Romani population p.c. have on average a 4.41 percentage lower number of communal workers per employed people, all other controls being equal. This association is significant at 5%. The second spline tells us that for observations having log Romani population p.c. larger than -6, observations with a ten percent higher Romani population p.c. have on average a 1.96 percentage higher number of communal workers per employed people, all other controls being equal. This association is significant at 1%. As noted earlier[10], the last coefficient is actually meaningless in reality on its own. What we can look at instead is the average of predicted values when the flag is one, meaning the initial value was zero. This tells us that the average predicted value of log communal workers per employed people is -3.82 when the flag is 1, or around 2.2 percent.

4. Robustness checks

Below, we present[11] two kinds of robustness checks: one for our models' performance on the test sample, and one for the sensitivity to the possible underreporting of the Romani population.

4.1. Model performance on test sample

To assess robustness, we take a look at some goodness of fit measures for the predicted values of the test sample (and we create these predicted values with the models built on the train sample). For the logit set-up, we can see that our chosen model performs worse on the test sample than on the training data. This indicates a possible overfitting issue in a technical sense. However, the magnitude of the increase in RMSE is acceptable (12.5 percent), so our model is rather robust. For the regular OLS model the goodness of fit measures indicate that the models actually fit the test sample slightly better than the train one. Thus, we can conclude that our OLS model is definitely not overfitted.

4.2. Sensitivity to the possible underreporting of Romani population

To assess the issue of underreporting, we generate new data with 5-25% increased values in Romani population p.c.[12], and re-estimate the same model specifications as before, to see how our models' parameter estimates compare to the models ran on the hypothetically corrected data. For the logit model, what we can see is that the marginal differences remain relatively stable despite the corrections applied. Thus our logit model is not sensitive to the underreporting of Romani population. For the OLS model, the coefficient of the spline below -6 changes substantially if a correction is applied. However, this comes at least partly from the fact that the larger the correction,

[9]Because of spatial constraints, we only interpret the results of the controlled models here.
[10]Recall that a 1% higher Romani population p.c. is meaningless if the actual baseline is zero.
[11]Here, we only present the interpretation of the robustness checks. For the detailed, numeric results, see Appx. A3.
[12]As we cannot know the true level of underreporting, if any.

the more observations switch their spline, thus the less observations remain below the knot. What is more important though is that the coefficient of the spline above -6 remains relatively stable despite the corrections. Thus we can say that this part of the model is not sensitive to the potential underreporting of Romani population.

5. Summary and room for causal interpretation

To sum up, we can see from the probability models that there is a significant positive association only if the Romani population p.c. is very small. The magnitude of this relationship is smaller as controls are included. As only a small fraction of observations fall into this lower spline, our probability models are rather inconclusive. The OLS models suggest that there is a significant negative association for the lower spline and a significant positive for the upper spline, with the upper spline's coefficient being smaller as controls are introduced. Both types of models perform relatively well on the test sample and are not sensitive to the possible underreporting of Romani population. Our results are only partly in line with our hypothesis - as coefficients diminish as controls are introduced, but significant relationships are still present in controlled models. This points towards the possibility that Romani people are overrepresented in marginalized social groups, but they still are positively discriminated[13] against in the context of communal work.

Note, however, that the above results cannot be interpreted as a causal relationship. This is because of two reasons. First, we only looked at a cross sectional sample, and we most probably could not include all relevant controls, so there may be some remaining uncontrolled heterogeneity between observations with different numbers of Romani population p.c. Second, the actual causal link that may be present is expected to happen on the individual level (that is Romani people may be more likely to be a communal worker just because they are of Romani ethnicity) - however, our data is aggregated on the municipal level, so we cannot make any causal claims without the possibility of committing an ecological fallacy. Finally, it is important to note that this study is based on observational, not experimental data. Experimental data would be necessary to establish causality by controlling for confounders.

6. Conclusion

Our hypothesis was that there is an initial positive association between the proportion of Romani individuals and the number of communal workers per employed person, but that this relationship would diminish as controls were introduced. In the logit model, the average predicted probability of having communal workers was high (80.8%) when the Romani ratio was 0, leaving limited room for explanatory power. One of the coefficients stayed statistically significant, though the association was weak and the other coefficient is not significant at all. In the OLS model, as expected, the association weakened by including more control variables, though it did not disappear entirely.

These results suggest that Romani people may be overrepresented in marginalized social groups, as the coefficients got smaller as controls were introduced to the models. The results also suggest that municipalities with higher numbers of Romani people p.c. have a higher number of communal workers per employed people, pointing towards the possibility of positive discrimination. With these results, we recommend policymakers to (1) create targeted programs for Romani communal workers that would help them reintegrate into the actual labor market; (2) conduct field studies and interviews with local communal work managers to examine the reason behind the possible overrepresentation of Romani people in communal work; and (3) create specific programs in marginalized municipalities to diminish the need for communal work.

As we mentioned earlier there are some limitations for our study. We examined municipality level data instead of individual data, which leaves open the possibility of an ecological fallacy, and it is also impossible to control perfectly for socio-economic and development indicators. Future research should explore the use of individual-level data to investigate these relationships more precisely.

[13]Positive only in the technical sense that they may be more likely to be communal workers.

Appendix

A1. Source data

A1.1. Source variables

The raw variables, as imported from TEIR and KSH, as well as their description can be found in Table A1. Note that some of these variables were not used directly in the analysis, but only for creating calculated ratio variables. Also note that variables related to the development of roads were dropped as missing value issues could not have been resolved.

Table A1: Description of the raw variables

Code	Short name	Description
muni_id	Municipality ID	ID of the municipality
muni_name	Municipality name	Name of the municipality
status	Status	Legal status of the municipality
county	County	County of the municipality
district	District	District (*járás*) of the municipality
comm_work	Communal workers	Number of communal workers (yearly average of monthly data)
romani_pctg	Romani percentage	Percentage of population belonging to Romani ethnicity.
dependence_rate	General dependency ratio	General dependency ratio (%)
pop	Population	Permanent population
pop_15_64f	Population (15-64 male)	15-64 year aged males from the permanent population
pop_15_64m	Population (15-64 female)	15-64 year aged females from the permanent population
county_center_min	Driving minutes to county center	Time to reach county center by road using the fastest route (minutes)
hprest_emp_pctg	High prest. employed percentage	Percentage of people employed in high-prestige occupations (% of employed people).
unemp_pctg	Unemployment percentage	Percentage of jobseekers registered for more than 180 days (%)
area	Area	Area of the municipality (km^2)
roads_undev	Undeveloped roads	Undeveloped municipality managed roads (km)
roads_dev	Developed roads	Developed municipality managed roads (km)
roads_total	Total roads	Total municipality managed roads (km)
income	Income	Consolidated income subject to personal income tax (HUF)
gp_visits	GP visits	Total yearly general practitioner visits
elem_students	Elementary students	Elementary students in full-time education
disadv_elem_students	Disadv. elementary students	Disadvantaged elementary students in full-time education
crime_per1000people	Crimes committed per 1K pop.	Registered crimes per 1000 inhabitants
low_comf_housing	Low comf. housing	Percentage of low-comfort houses and occupied holiday homes (%)
educ_level_max_elem	Max. elementary educ. pop.	Number of people with maximum elementary education in the 7-X population
migration_per1000	Migration per mille	Domestic migration balance, per thousand inhabitants
growth_decline_per1000	Natural pop. growth/decline per mille	Natural population growth or decline (per mille)
elem_stud_from_other_muni_pctg	Elem. stud. from other muni. percentage	Percentage of elementary school students in full-time education from another municipality (%)
pop_65-x_m	Population (65-X male)	65-X year aged males from the permanent population
pop_65-x_f	Population (65-X female)	65-X year aged females from the permanent population
people_in_agriculture	Agric. employed	Number of people employed in agriculture (FEOR-08: 61)
employed	Employed	Number of employed people
pop_7_x	Population (7-X)	7-X year aged population from the permanent population

A1.2. Calculated variables and rescaling

For most of our variables that were not ratios by definition, we created some ratio variables by dividing with a sensible reference group. These included the following:

- Population density (*pop_dens*): we divided the permanent population by the area of the municipality;
- Old age dependency ratio (*old_age_dependency_ratio*): ratio of the sum of population (65-X) male and female and the sum of population (15-64) male and female;
- Communal workers per employed people (*comm_work_per_employed*): ratio of communal workers and employed people;
- Monthly income per capita (*monthly_income_pc*): income divided by population[14] and 12;
- Maximum elementary education ratio (*max_elem_educ_ratio*): ratio of population with maximum elementary education and 7-X aged population;
- Agriculture employed ratio (*agr_emp_ratio*): ratio of people employed in agriculture and total number of employed people;

[14]Note that as we divide by population here, the resulting variable does not capture average income in the usual sense (that is income of actual earners), but only in a per capita sense. Also note that this is income before taxation.

- Disadvantaged elementary students ratio (*disadv_elem_students_ratio*): ratio of disadvantaged full time elementary students and total number of full time elementary students;
- Average GP visits per capita (*avg_gp_visits_pc*): the total number of GP visits divided by the permanent population;
- Elementary students from other municipality ratio (*elem_stud_from_other_muni_ratio*): the number of full time elementary students from other municipalities divided by the total number of full time elementary students.

Apart from these variables, we also created two dummy variables for the municipality status, with the reference group being villages (*község* in Hungarian): one for towns (*nagyközség* in Hungarian) and one for cities (everything other than villages and towns). This set-up seemed more sensible than including a dummy for every possible category, as larger city categories (e.g. *megyei jogú város*) naturally had very few observations.

Lastly, we scaled all percentage variables down to a 0 to 1 scale by dividing with a 100. This allows us to interpret most of these variables (and especially the ratio of Romani population) in a sort of per capita sense (e.g. Romani population per capita), which may be more intuitive if we happen to work with log transformed variables[15].

A1.3. Summary statistics and distributions

We present the summary statistics for the continuous variables (after the imputations described in Sec. 2.2.) we plan to use in our analysis[16] in Table A2. Note that from this point on, the results presented relate only to the train sample.

Table A2: Summary statistics of continuous, non-transformed variables

Variable	Count	Mean	Std.	Min.	5%	25%	50%	75%	95%	Max.
Communal work. per emp.	2,522	0.059	0.088	0.000	0.000	0.007	0.025	0.076	0.229	0.972
Romani population p.c.	2,522	0.044	0.084	0.000	0.000	0.000	0.012	0.048	0.207	0.902
Monthly income p.c.	2,522	143,894	44,442	23,215	76,550	112,905	141,648	171,671	219,872	490,138
Unemployment ratio	2,522	0.520	0.195	0.000	0.158	0.421	0.529	0.636	0.800	1.000
High. prest. employed ratio	2,522	0.157	0.085	0.000	0.055	0.101	0.137	0.193	0.320	0.714
Agric. employed ratio	2,522	0.012	0.017	0.000	0.000	0.001	0.006	0.015	0.043	0.176
Population density	2,522	72.699	125.369	1.515	11.416	26.135	42.996	73.109	211.151	2094.408
Old age dependency ratio	2,522	0.273	0.088	0.037	0.157	0.221	0.261	0.314	0.432	1.085
General dependency ratio	2,522	0.547	0.091	0.115	0.420	0.495	0.541	0.591	0.694	1.250
Migration per mille	2,522	0.546	34.584	-514.290	-47.998	-12.875	0.565	15.230	44.859	307.290
Nat. pop. growth/ decline per mille	2,522	-6.333	11.659	-99.070	-25.196	-10.608	-5.115	0.000	8.888	40.070
Max. elem. educ. ratio	2,522	0.278	0.083	0.000	0.157	0.218	0.270	0.327	0.421	0.645
Disadv. elem. stud. ratio	2,522	0.064	0.118	0.000	0.000	0.000	0.000	0.077	0.326	0.920
Elem stud. from other muni. ratio	2,522	0.135	0.209	0.000	0.000	0.000	0.000	0.213	0.612	1.000
Driving minutes to county center	2,522	44.092	20.497	0.000	15.571	29.130	41.705	55.990	80.799	144.940
Low comf . housing ratio	2,522	0.059	0.085	0.000	0.000	0.000	0.019	0.095	0.229	0.621
Crime committed per 1K pop.	2,522	13.710	13.750	0.000	0.000	6.510	10.720	17.022	34.596	244.330
Avg. GP visits p.c.	2,522	2.305	0.362	0.604	1.759	2.076	2.285	2.524	2.914	3.877
Population	2,522	2408	6,954	17	109	353	828	1,992	8,753	141,761

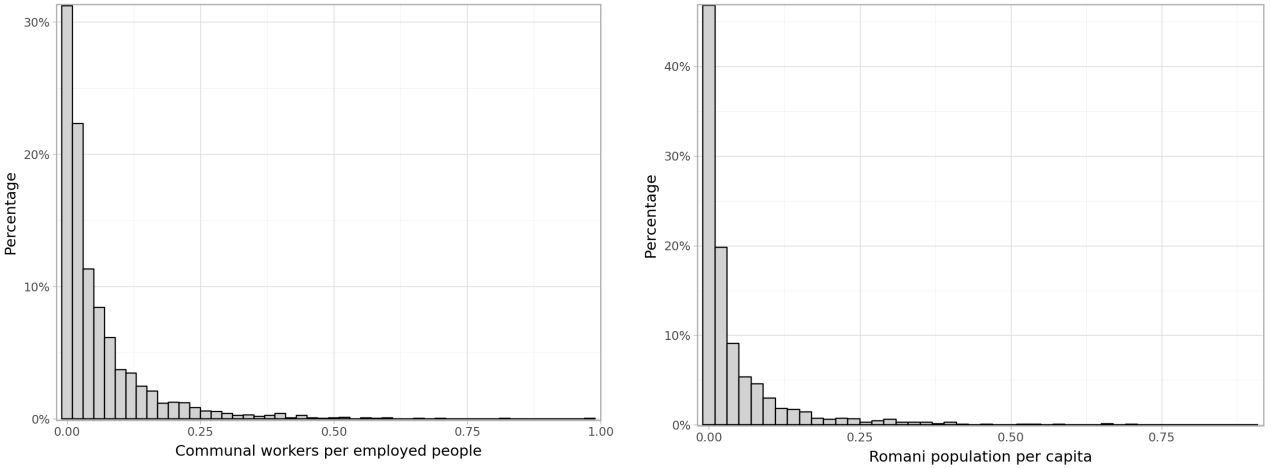
As we can see from the above table, our two main variables of interest are rather right-skewed. This is even more apparent from the histograms[17] presented in Figure A1.

[15]E.g. it is more understandable to talk about percentage changes in Romani population per capita than a percentage change in the ratio of Romani population, let alone a percentage change in the percentage of Romani population.

[16]Thus, those variables which were only used to create ratio variables are not presented here.

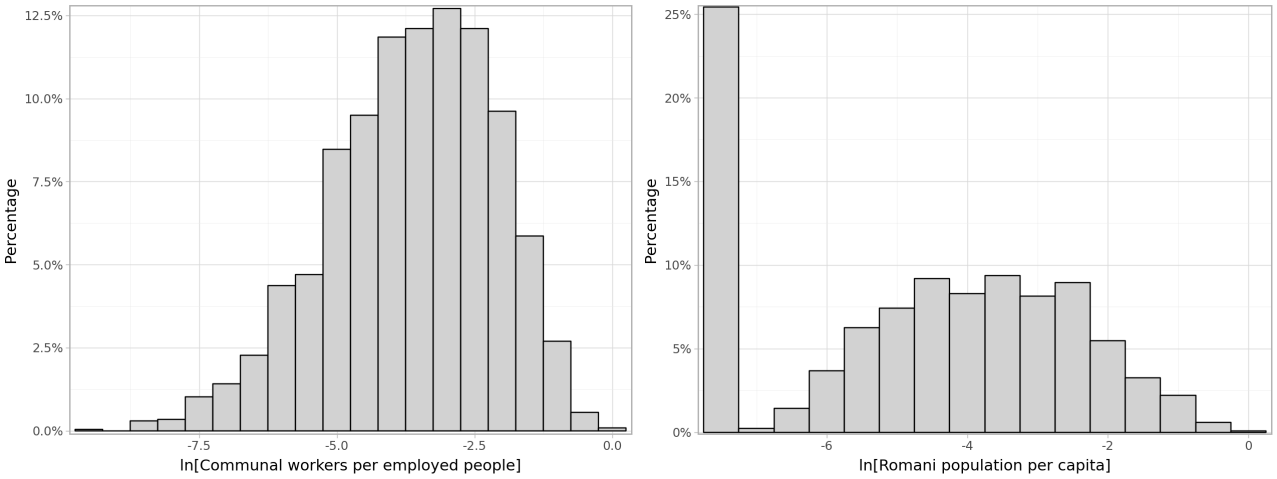
[17]Even though we have checked the histograms of all other variables as well, these are not presented here as they are not crucial to understand our results and strain of thought. This remark is also true for lowess estimates.

Figure A1: Histograms of the non-transformed main variables (N=2,522)



Because of the many zero-values, we applied a log transformation as described in Sec. 2.2. The resulting histograms are presented in Figure A2.

Figure A2: Histograms of the log-transformed main variables (N=2,338 and 2,522)



Note: for the log communal workers, only the observations with non-zero initial values (N=2,338) are presented.

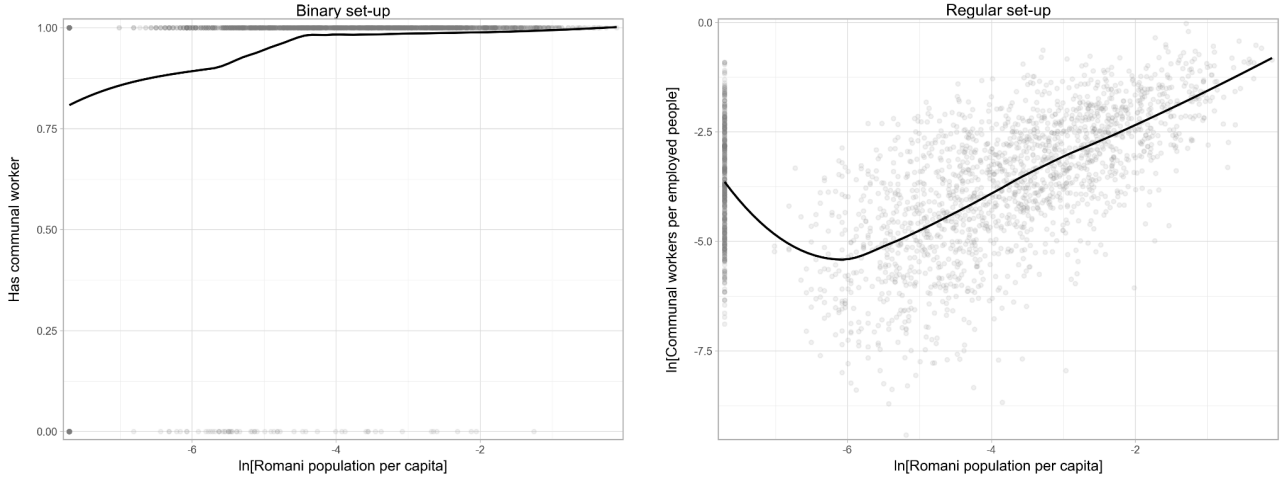
As we can see, the distribution of log Romani population per capita is still far from normal - but this will be handled by including the flag variable in our analysis.

In addition to our main variables, we also applied a log transformation to the following control variables because of their positive minima and skewed distribution: monthly income per capita, population density, old age and general dependency ratio, and population. All other controls were kept at level.

A1.4. Lowess estimates for the main variables

To uncover the possible functional forms between our binary and regular outcome variables and Romani population per capita, we estimate two lowess regressions. These are presented in Figure A3. What we can see is that a linear spline specification may be a good option with knots at -4.5 and -6 for the binary and regular set-ups, respectively.

Figure A3: Lowess estimates for log Romani population per capita



A2. Regression analysis

Before moving on to the actual analysis, let us note some additional data issues we encountered during our analysis. Most importantly, we noticed that our flag variables could not be included in any models the way regular dummy variables can (that is, on their own and as an interaction). The reason behind this was perfect collinearity. For the Romani flag, this meant that for each observation the flag was 1, ln[Romani population p.c.] was the same minimal value we assigned to it in the beginning. So the interaction term became simply that same value for every observation where the flag was one - and thus, this contained exactly the same information as the flag on its own. For better interpretation, we decided to include the interaction term only.

For the no elementary students flag, its value was 1 for observations where the elementary students from other municipality ratio was zero - thus an interaction term would have become only zeros, which does not contain any information. So we could only include this flag on its own.

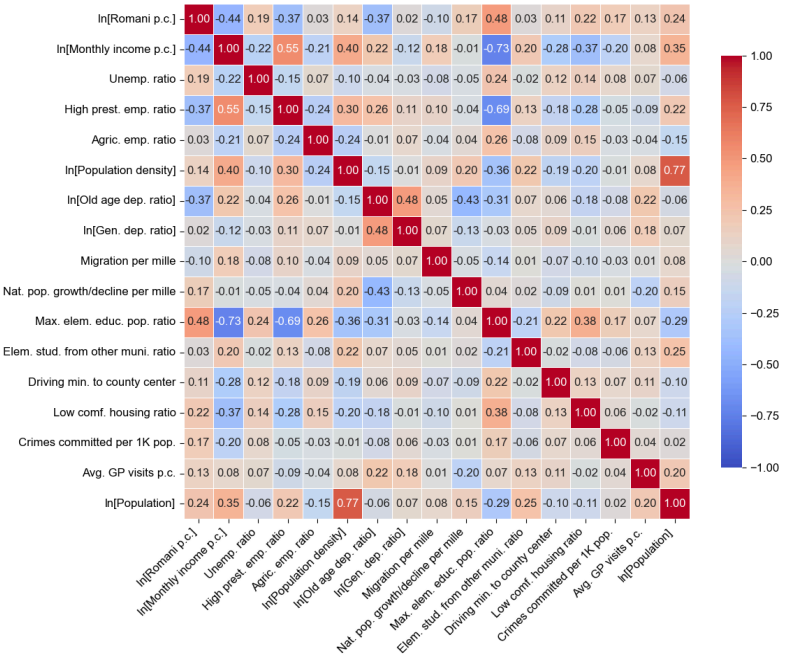
Other than the flag variables, we decided to exclude from the analysis the disadvantaged elementary students ratio, as it was causing quasi-complete separation in our logit models, and strong collinearity issues in our regular OLS models, which we could not resolve with principal components. We also excluded the area of the municipality, as all of its information content was already captured by the population and the population density.

Lastly, we could not include county dummies in the logit models, as there was practically no within-county variation in the has communal worker binary, leading to a singular matrix and an inestimable model. However, this was not an issue for the regular OLS models, as our dependent variable was different.

A2.1. Collinearity analysis and principal components

When examining our variables in detail, we noticed that many of them serve as proxies for the developmental level or socio-economic status of a municipality. This can pose a challenge if the inclusion of multiple proxies introduces multicollinearity into our models. To assess this, we visualize the correlations among the non-binary explanatory variables in Figure A4.

Figure A4: Heatmap of the correlation matrix



As expected, some variables exhibit strong correlations, especially high prestige employment ratio and monthly income per capita or high prestige employment ratio and maximum elementary education ratio. To address this issue, we turned to principal component analysis (PCA), a dimensionality reduction technique. While we could have calculated the averages of z-scores within specific variable groups, we chose not to do so for two reasons. First, even z-score averages are challenging to interpret (as the variables have different scales), so we lose interpretability either way. Second, if we are losing interpretability, PCA provides a more accurate and robust approach than simple z-score averaging.

PCA requires scaling the variables beforehand, so we ensured that all variables are properly standardized. Initially, we created five principal component variables based on our intuition. However, these components proved difficult to interpret, as they combine variables where higher values can have conflicting implications (e.g., higher is better for one variable but worse for another). To address this, we later created more refined and meaningful principal components, presented in Table A3. Importantly, we generated the principal components for the test sample using the weights derived from the training sample, ensuring that we do not inadvertently use information from the test sample. For each variable group, we used only the first principal component. All other variables not included in principal components were included on their own in the models.

Table A3: Variables included in principal components

Principal component's name	Variables included
Health & safety	low comf. housing ratio, crimes committed per 1K pop., avg. GP visits p.c.
Migration & mobility	migration per mille, natural pop. growth/decline per mille
Job composition (positive)	ln[monthly income p.c.], high prest. employed ratio
Job composition (negative)	unemployment ratio, agric. employed ratio
Demography	ln[old age dependency ratio], ln[general dependency ratio]
Population & size	ln[population], ln[population density]

A2.2. Estimated probability models

We estimated 10 different probability models in total:

- Model 1: includes only one explanatory variable - the log Romani p.c. - in a linear form. As expected, the relationship is positive and significant.
- Model 2: includes the Romani p.c. in a logarithmic form. Since the Romani p.c. can take a value of zero, a flag variable is added for these observations.
- Model 3: Features the log Romani p.c. modeled with a piecewise linear spline, with a knot at -4.5.

- Model 4: includes more explanatory variables but Romani p.c. stays in a linear way.
- Model 5: includes log Romani p.c. with piecewise linear spline with a knot of -4.5 but includes several explanatory variables in linear or logarithmic forms.
- Model 6: includes log Romani p.c. with piecewise linear spline with a knot of -4.5 and principal components with straightforward positive or negative association.
- Model 7: Expands on Model 6 by allowing for non-linear patterns in the principal components.
- Model 8 is different from model 7 with only including the significant explanatory variables[18].
- Model 9 doesn't include principal components, only the explanatory variables, but it allows non-linear patterns if necessary.
- Model 10 is different from model 9 because it only includes significant variables.

The results of these 10 models are presented in Table A4.

Table A4: Logit regression results - all estimated models' marginal differences

(dependent variable: has communal worker binary; N=2,522)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Romani p.c.	2.627***			0.413						
	(0.894)			(0.258)						
ln[Romani p.c.]		0.042***								
		(0.008)								
ln[Romani p.c.] x flag		-0.003								
		(0.003)								
ln[Romani p.c.] spline < -4.5			0.055***		0.023*	0.028**	0.027**	0.026**	0.026**	0.023**
			(0.012)		(0.013)	(0.012)	(0.012)	(0.011)	(0.012)	(0.011)
ln[Romani p.c.] spline > -4.5			0.029**		-0.007	-0.006	-0.010	-0.007	-0.009	-0.002
			(0.013)		(0.012)	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)
ln[Romani p.c.] spline < -4.5 x flag			-0.007*		-0.000	-0.000	-0.001	-0.001	-0.002	-0.001
			(0.004)		(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Control variables										
Unemployment ratio				0.113***	0.103***				Yes***	Yes***
				(0.023)	(0.022)				(poly.)	(poly.)
High prest. employed ratio				-0.172**	-0.170**				-0.223***	-0.263***
				(0.069)	(0.067)				(0.076)	(0.067)
Agric. employed ratio				0.020	0.092					
				(0.366)	(0.358)					
Migration per mille				0.000	0.000				Yes**	Yes**
				(0.000)	(0.000)				(lq. spline)	(lq. spline)
Natural pop. growth /decline per mille				-0.000	-0.000				Yes	
				(0.001)	(0.000)				(lin. spline)	
Max. elemenetary educ. pop. ratio				0.394***	0.333***	0.341***	Yes***	Yes***	Yes***	Yes***
				(0.107)	(0.101)	(0.100)	(lin.spline)	(lin.spline)	(lin.spline)	(lin.spline)
Elem. students from other muni. ratio				0.085**	0.083**	0.068*	0.061*	0.059*	0.070**	0.061**
				(0.038)	(0.037)	(0.036)	(0.034)	(0.032)	(0.033)	(0.031)
Driving minutes to county center				-0.000	-0.000	-0.000	-0.000		-0.000	
				(0.000)	(0.000)	(0.000)	(0.000)		(0.000)	
Ratio of low comf. housing				0.168*	0.168*				Yes**	Yes*
				(0.099)	(0.096)				(lin. spline)	(lin. spline)
Crimes commited per 1K pop.				0.000	0.000					
				(0.000)	(0.000)					
Avg. GP visits p.c.				0.046***	0.043***				Yes**	Yes*
				(0.016)	(0.015)				(lin. spline)	(lin. spline)
ln[Monthly income p.c.]				-0.062**	-0.053*				Yes (poly.)	
				(0.030)	(0.029)					
ln[Population density]				-0.019**	-0.019**				Yes**	Yes**
				(0.009)	(0.008)				(poly.)	(poly.)
ln[Old age dependency ratio]				-0.050**	-0.048*				Yes (lin. spline)	
				(0.025)	(0.025)					
ln[General dependency ratio]				0.002	0.008				Yes**	Yes***
				(0.042)	(0.040)				(poly.)	(poly.)
ln[Population]				0.038***	0.028***				0.010	
				(0.009)	(0.009)				(0.010)	
Health & safety PC						0.016**	Yes (lin. spline)			
						(0.007)				
Migration & mobility PC						-0.004	Yes**	Yes**		
						(0.005)	(poly.)	(poly.)		
Demography PC						-0.007	Yes (lin. spline)			
						(0.004)				

[18]As this is rather a proof-of-concept study, we decided to keep controls that were significant at 10%.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Population & size PC						0.007 (0.006)	Yes (poly.)			
Job composition (positive) PC						-0.024*** (0.006)	Yes*** (lin.spline)	Yes*** (lin.spline)		
Job composition (negative) PC						0.024*** (0.006)	Yes*** (lq. spline)	Yes*** (lq. spline)		
No elem. students dummy				0.001 (0.020)	0.009 (0.020)	-0.018 (0.019)	-0.020 (0.018)	-0.014 (0.016)	-0.000 (0.020)	-0.008 (0.016)
Status dummies				Yes	Yes	Yes	Yes		Yes	
R-squared	0.055	0.081	0.080	0.191	0.206	0.190	0.224	0.213	0.260	0.250
Brier-score	0.064	0.062	0.062	0.055	0.054	0.055	0.052	0.053	0.050	0.051
Pseudo R-squared	0.090	0.148	0.149	0.281	0.296	0.274	0.313	0.307	0.343	0.333
Log-loss	-0.238	-0.223	-0.222	-0.188	-0.184	-0.190	-0.180	-0.181	-0.172	-0.174

Note: Heteroskedasticity robust standard errors (HC1) in parentheses, if applicable . For control variables, standard errors are only reported if the variable is included in linear form and if it is not a categorical variable. For non-linear and categorical controls, their inclusion is indicated by a Yes *label and the number of significance stars corresponding to the most significant non-linear part. The functional form is reported in parentheses (lin. spline - piecewise linear spline; lq. spline - linear spline below knot, quadratic spline above; poly. - polynomial form). *p<0.1; **p<0.05; ***p<0.01*

As we can see, most of our goodness of fit measures tend to agree that the best PC model would be model (7) and the best non-PC model would be model (9). However, these models are a bit overfitted in the sense that they contain every possible variable with no regards to the significance level. A slightly better option would be in our opinion to use models (8) and (10) instead, as they contain only the variables that were significant in the previous models. To decide which model is better from these two, we plot two diagnostic charts, and also calculate confusion tables, as presented in A2.3.

A2.3. Probability model diagnostics

First, we examine the distribution of predicted values conditional on the actual value of *has communal worker* for our two chosen models in Figure A5. These offer no clean-cut decision on which model would be better. The overlap between the distributions is rather similar.

Figure A5: Distribution of predicted values conditional on the actual value for the two best performing probability model

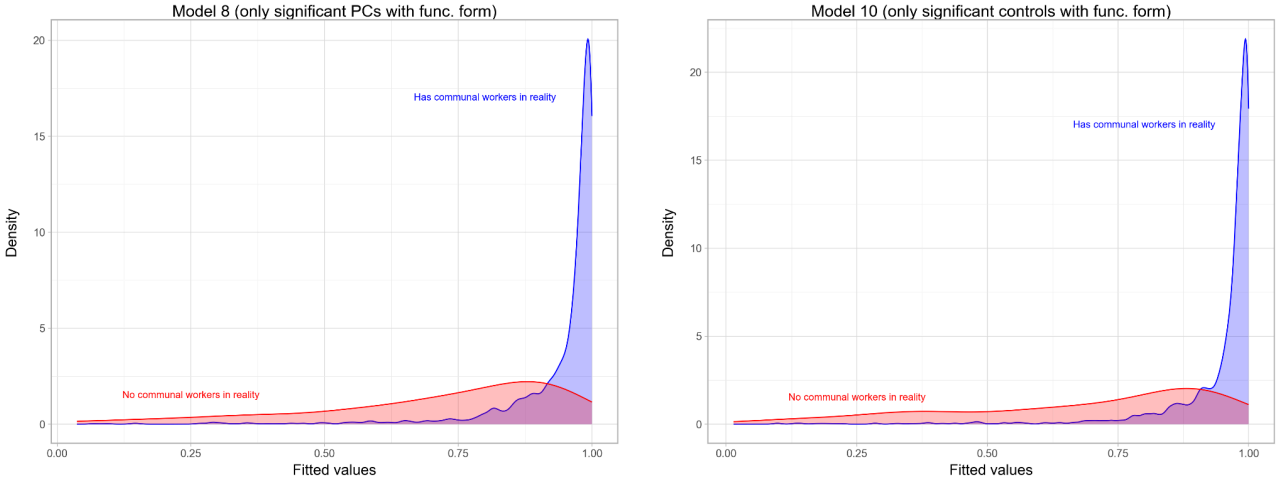


Table A5 shows the confusion rate of the two models with a cut-off at 0.75[19]. Model 10 identifies the true values slightly better with a confusion rate of 7.34% as compared to Model 8's 8.01%. However, the difference in confusion rates is relatively small, so the choice between models might depend on other factors.

[19]Note that this cut-off rate has been decided rather arbitrarily and different cut-offs would result in different confusion rates.

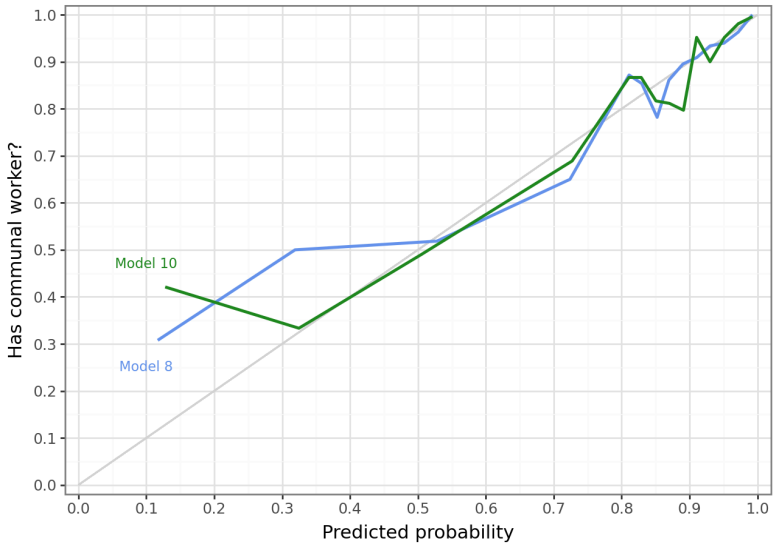
Table A5: Confusion table for the two best performing probability model

		Model 8 predictions for *has communal worker*		Model 10 predictions for *has communal worker*	
		no	**yes**	**no**	**yes**
Actual values for *has communal worker*	**no**	83	101	90	94
	yes	101	2,237	91	2,247
Total confusion rate		8.01%		7.34%	

Note: cut-off for classification at 0.75

The last diagnostic plot we look at is a calibration plot for the two models (Figure A6)[20]. What we can see from this is that there is a trade-off between the models: Model 10 is better calibrated for lower probabilities, while Model 8 is better calibrated for higher ones. As most of our sample has communal workers, we think that better calibration is more important for higher probabilities, thus we opt for Model 8 as our chosen model.

Figure A6: Calibration plot for the two best performing probability model



Note: 20 percentage point bins below 80%, 2 percentage point bins above 80%

A2.4. Estimated regular OLS models

Similarly to the probability models, we estimated 10 regular OLS models in total. The idea is the same behind the models:

- Models (1)-(3) aim to uncover a reasonable functional form for the Romani population per capita. Just like before, we settle for a piecewise linear spline of the log transformed variable. Note that in Model (4) we experimented with leaving the Romani population per capita at level while including all controls.
- Model (5) (and all further models) includes the log Romani population per capita with the flag interaction, and it controls for all raw controls (that is, not principal components) in linear form.
- Model (6) controls for all principal components in linear form, and it also includes those raw controls that were not included in any principal components.
- Model (7) adds functional forms to the principal components.
- Model (8) keeps only the significant[21] controls from the previous model.
- Model (9) controls for all raw controls with functional forms added.
- Model (10) keeps only the significant controls from the previous model.

All the above models are presented in detail in Table A6.

[20]We have also calculated the overall bias of the models which showed that both of them were practically unbiased.
[21]As this is rather a proof-of-concept study, we decided to keep controls that were significant at 10%.

Table A6: Regular OLS results - all estimated models

(dependent variable: ln[communal workers per employed people]; N=2,338)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Romani p.c.	8.035***			0.948***						
	(0.411)			(0.229)						
ln[Romani p.c.]		0.781***								
		(0.019)								
ln[Romani p.c.] x flag		-0.409***								
		(0.014)								
ln[Romani p.c.] spline < -6			-0.369		-0.405*	-0.543**	-0.442**	-0.441**	-0.405*	-0.404*
			(0.312)		(0.222)	(0.221)	(0.213)	(0.212)	(0.224)	(0.222)
ln[Romani p.c.] spline > -6			0.802***		0.191***	0.235***	0.196***	0.196***	0.182***	0.179***
			(0.020)		(0.018)	(0.018)	(0.017)	(0.017)	(0.018)	(0.018)
ln[Romani p.c.] spline < -6 x flag			-0.163**		0.049	0.066	0.050	0.050	0.050	0.052
			(0.067)		(0.048)	(0.048)	(0.046)	(0.046)	(0.049)	(0.048)
Constant	-4.023***	-0.774***	-7.737***	6.971***	3.133*	-7.490***	-8.072***	-8.082***	-76.905***	-78.155***
	(0.035)	(0.065)	(1.895)	(1.327)	(1.878)	(1.345)	(1.323)	(1.318)	(15.882)	(16.091)
Control variables										
Unemployment ratio				0.188**	0.157*				0.167**	0.175**
				(0.087)	(0.083)				(0.084)	(0.084)
High prest. employed ratio				-1.967***	-1.746***				Yes***	Yes***
				(0.293)	(0.290)				(poly.)	(poly.)
Agric. employed ratio				-1.023	-0.837				-1.043	
				(0.792)	(0.782)				(0.746)	
Migration per mille				-0.001	-0.001				Yes*	Yes*
				(0.000)	(0.000)				(lin. spline)	(lin. spline)
Nat. pop. growth/ decline per mille				0.001	0.000				Yes	
				(0.002)	(0.001)					
Max. elemenetary educ. pop. ratio				3.045***	2.563***		2.482***	2.496***	2.343***	2.332***
				(0.331)	(0.314)		(0.351)	(0.345)	(0.328)	(0.328)
Elem. stud. from other muni. ratio				0.020	0.073	0.084	0.196**	0.198**	0.112	
				(0.090)	(0.088)	(0.087)	(0.086)	(0.086)	(0.088)	
Driving minutes to county center				0.005***	0.005***	0.004***	0.004***	0.005***	0.005***	0.005***
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Ratio of low comf. housing				0.363**	0.395**				0.458***	0.456***
				(0.161)	(0.155)				(0.152)	(0.153)
Crimes commited per 1K pop.				0.002**	0.002*				0.001	
				(0.001)	(0.001)				(0.001)	
Avg. GP visits p.c.				0.110**	0.077				Yes**	Yes***
				(0.055)	(0.053)				(lin. spline)	(lin. spline)
ln[Monthly income p.c.]				-0.736***	-0.625***				Yes***	Yes***
				(0.114)	(0.107)				(poly.)	(poly.)
ln[Population density]				-0.149***	-0.142***				-0.134***	-0.135***
				(0.030)	(0.029)				(0.029)	(0.029)
ln[Old age dependency ratio]				-0.145*	-0.083				Yes**	Yes**
				(0.077)	(0.069)				(lin. spline)	(lin. spline)
ln[General dependency ratio]				0.264**	0.238**				Yes**	Yes**
				(0.129)	(0.121)				(lin. spline)	(lin. spline)
ln[Population]				-0.359***	-0.371***				-0.359***	-0.327***
				(0.031)	(0.032)				(0.034)	(0.028)
Health & safety PC						0.058***	Yes***	Yes***		
						(0.015)	(lin. spline)	(lin. spline)		
Migration & mobility PC						0.020	Yes			
						(0.017)	(lq. spline)			
Demography PC						0.005	Yes			
						(0.016)	(lin. spline)			
Population & size PC						-0.400***	-0.347***	-0.350***		
						(0.023)	(0.023)	(0.023)		
Job composition (positive) PC						-0.368***	Yes***	Yes***		
						(0.026)	(poly.)	(poly.)		
Job composition (negative) PC						0.014	Yes*	Yes*		
						(0.016)	(lq. spline)	(lq. spline)		
No elem. students dummy				-0.102*	-0.104*	-0.018	0.030	0.028	-0.081	
				(0.055)	(0.054)	(0.048)	(0.047)	(0.047)	(0.053)	
Status dummies				Yes***	Yes***	Yes***	Yes***	Yes***	Yes***	Yes***
County dummies				Yes***	Yes***	Yes***	Yes***	Yes***	Yes***	Yes***
R-squared	0.212	0.394	0.396	0.812	0.820	0.809	0.817	0.817	0.824	0.823
Adj. R-squared	0.211	0.393	0.395	0.808	0.817	0.807	0.814	0.814	0.821	0.820
BIC	8016.1	7410.1	7408.5	4957.1	4865.7	4937.7	4907.3	4872.1	4866.1	4834.3
AIC	8004.5	7392.8	7385.4	4732.5	4629.7	4747.7	4665.5	4659.1	4589.8	4592.5

Note: Heteroskedasticity robust standard errors (HC1) in parentheses, if applicable . For control variables, standard errors are only reported if the variable is included in linear form and if it is not a categorical variable. For non-linear and categorical controls, their inclusion is indicated by a Yes *label and the number of significance stars corresponding to the most significant non-linear part. The functional form is reported in parentheses (lin. spline -*

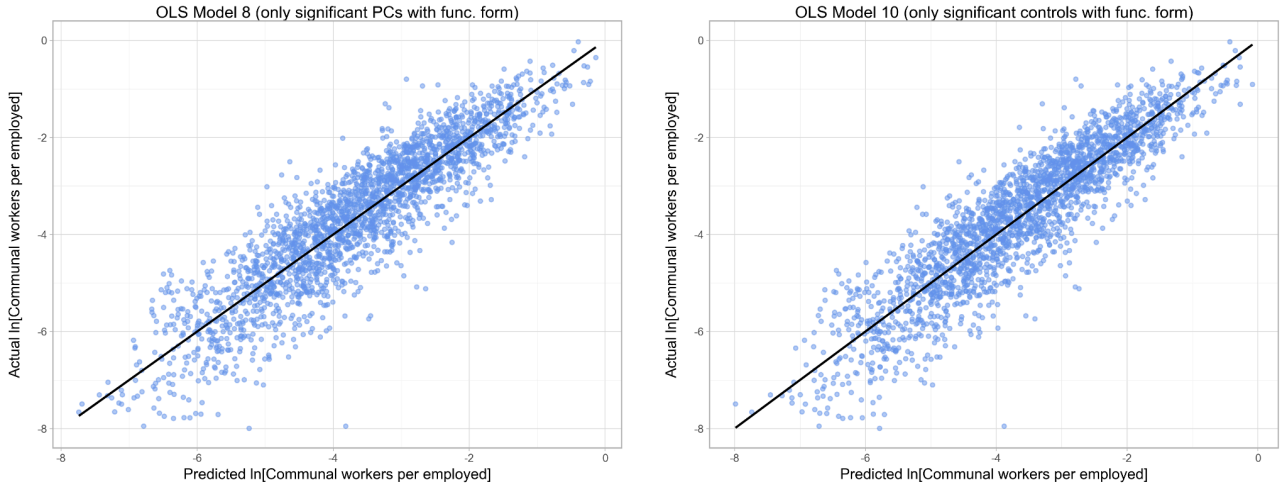
*piecewise linear spline; lq. spline - linear spline below knot, quadratic spline above; poly. - polynomial form). *p<0.1; **p<0.05; ***p<0.01*

As we can see from the goodness of fit indicators and the information criteria, excluding non-significant controls from models (7) and (9) does not entail a meaningful loss of precision. The resulting models (8) and (10) seem rather similar in their goodness of fit, so we will examine these models further in the next section.

A2.5. Regular OLS model diagnostics

As we can see from the ŷ-y plots presented in Figure A7, the two best performing models produce rather similar predictions. Thus, we decided to deem Model (8) the best again. This way, we traded a tiny bit of predictive power to mitigate the possibility of multicollinearity and inflated standard errors that may be present in Model (10).

Figure A7: ŷ-y plots of the two best performing OLS model



A3. Robustness checks

As described in Sec. 4, we present two kinds of robustness checks: one for our models' performance on the test sample, and one for the possible underreporting of the Romani population in our source data.

A3.1. Probability model fit on the test sample

From Table A7, we can see that our logit model performs worse on the test sample than on the training data (as indicated by an increase in MSE, RMSE and log-loss, and a decrease in R-squared). This might indicate a possible overfitting issue. However, if we take a look at the RMSE values, what we can see is that in our original model, we were 23.1 percentage points off on average in our predictions, which increased to 26.0 percentage points in the test sample (a 12.5% increase). So actually, the magnitude of the increase is rather tolerable, and our model performs relatively well on unseen data.

Table A7: Probability model results on the test sample

Sample	R-squared	MSE	RMSE	Log-loss
Train	0.214	0.053	0.231	-0.181
Test	0.165	0.068	0.260	-0.240

A3.2. Regular OLS model fit on the test sample

From Table A8, we can see that all of the measures indicate that our regular OLS model fits even better to the test sample than to the train (higher R-squared and adjusted R-squared and lower MSE and RMSE). Though this is a rather unlikely result, this may just come from the random way we have split our dataset into a train and test sample. So we can conclude that our regular OLS model is definitely not overfitted to the training data.

Table A8: Regular OLS model results on the test sample

Sample	R-squared	Adj. R-squared	MSE	RMSE
Train	0.817	0.814	0.416	0.645
Test	0.836	0.825	0.380	0.616

A3.3. Probability model sensitivity to Romani population underreporting

We might be concerned about the fact that ethnicities may be systematically underreported in census data, especially if a certain ethnicity is stigmatized. Even though we cannot really check whether this is the case, we can test our models' stability under different hypothetical scenarios. For this, we generated new data with 5-25% increased values in Romani ratio, and re-estimated the same model specifications as before, to assess how our models' parameter estimates compared to the models ran on the hypothetically corrected data.

What we can see from Table A9 is that the marginal differences remain relatively stable despite the corrections applied (or they were insignificant, and remained insignificant). Thus our logit model is not sensitive to the potential underreporting of the Romani population.

Table A9: Probability model sensitivity to Romani population underreporting (marginal differences, N=2,522)

	Original results	**+5% correction**	**+10% correction**	**+15% correction**	**+20% correction**	**+25% correction**
ln[Romani p.c.]	0.026**	0.026**	0.026**	0.026**	0.026**	0.025**
spline < -4.5	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)
ln[Romani p.c.]	-0.007	-0.006	-0.005	-0.004	-0.003	-0.002
spline > -4.5	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
ln[Romani p.c.]	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
spline < -4.5 x flag	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)

*Note: Heteroskedasticity robust standard errors (HC1) in parentheses. *p<0.1; **p<0.05; ***p<0.01*

A3.4. Regular OLS model sensitivity to Romani population underreporting

What we can infer from the results in Table A10 is that the coefficient of the spline below -6 changes substantially if a correction is applied for underreporting. However, this comes at least partly from the fact that the larger the correction, the more and more observations switch their spline, thus the less observations remain below the knot. What is more important though is that the coefficient of the spline above -6 remains relatively stable despite the corrections. Thus we can say that at least this part of the model (where most of our sample is) is not sensitive to the potential underreporting of the Romani population.

Table A10: Regular OLS model sensitivity to Romani population underreporting (N=2,338)

	Original results	**+5% correction**	**+10% correction**	**+15% correction**	**+20% correction**	**+25% correction**
ln[Romani p.c.	-0.441**	-0.504**	-0.572**	-0.646**	-0.735***	-0.843***
spline < -6	(0.212)	(0.229)	(0.246)	(0.262)	(0.277)	(0.289)
ln[Romani p.c.]	0.196***	0.195***	0.194***	0.194***	0.193***	0.193***
spline > -6	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
ln[Romani p.c.]	0.050	0.063	0.077	0.092	0.111*	0.134**
spline < -6 x flag	(0.046)	(0.050)	(0.054)	(0.057)	(0.061)	(0.063)

*Note: Heteroskedasticity robust standard errors (HC1) in parentheses. *p<0.1; **p<0.05; ***p<0.01*