University of Cape Town

STA5069Z

Multivariate Statistics

Perceived Discrimination in the European Labor Market: Demographic Determinants and Cross-Country Variations

Author:
Xiaojing Huang

Student Number:
HNGXIA002

April 5, 2025

Contents

1 Introduction

The participation of diverse labour force is crucial for the economic growth in a country as a wider range of perspectives, skills, and experience can be brought into the country. Migrants, who are also part of the labour market, often face unique challenges in their hosting countries, including legal barriers and language proficiency. According to statistics from international migration stock data (2024), nearly 87 million international migrants lived in Europe. Of the 87 million migrants, there were around 44 million who were born within Europe but lived elsewhere in the region, and around 40 millions of non-European migrants resided in European regions (International Organization of Migration, 2020). Despite their significant presence, migrant workers experience sizeable employment gaps compared to native workers.

The study by Giang Ho and Rima Turk-Ariss on the labour market integration of migrants in Europe (2018) found that the employment opportunities for migrants would gradually converge to those of natives, but full coverage has not been observed even after 20 years. The persistent employment gap highlights the complexity of migration integration and suggests that discrimination, along with other structural barriers, may continue to affect the experience of migrants on the job. Another interesting perspective to investigate is the so-called "Integration paradox". The integration paradox suggests that higher educated migrants report that they have experienced higher level of discrimination than those who are less integrated (Verkuyten, 2016). Furthermore, international migration trend shows that nowadays almost as many females as males migrate to and within Europe, and females follow the new trend of migrating on their own for the purpose of searching for jobs, in contrast to the purpose of family unification as in the past (Cortinovis et al., 2020).

Therefore, this study aims to use Multinomial Logistic Regression to first evaluate the significance and influence of demographic backgrounds of migrants on the response pattern and whether the model fit has predictive power on predicting the likelihood of perceiving discrimination. Specifically, the study will examine the gender level of inequality in terms of perceiving discrimination whether highly skilled migrants are more likely to over-report or under-report their discrimination perception in order to provide further evidence to the integration paradox. After exploring the relationships among factors, the study will use Latent Class Analysis to identify distinct subgroups of migrants who perceive discrimination differently based on their socio-demographic characteristics. The goal of this study is to base on these results to provide insightful practical recommendations that can help address such discrimination situation in Europe.

2 Hypotheses

Despite the statistics show that there are almost as many female as male migrants in Europ, as the female migrants first enter the hosting countries, their proficiency in hosting-country language and familiarity in social norms impose a barrier for them to find the jobs in those fields (Schieckoff, Sprengholz, 2021; Das, Kotikula, 2019; Raijman, Semyonov, 1997). Therefore, below is the first hypothesis related to perceived gender discrimination:

H1:Female migrants are more likely to perceive discrimination compared to male migrants.

The argument that supports the integration paradox suggests that highly educated migrants have more social exposure to mainstream society and, therefore, have more opportunity to face discrimination (Schaeffer, Kas, 2023). Therefore, it is hypothesized that:

H2: Migrants who have received tertiary education are more likely to over-report discrimination compared to migrants who are less educated.

Despite that EU countries have agreed to use certain common immigration and permit rules, there are other aspects which each country can develop their own rules (European Commission, 2025). This implies that immigration patterns are not identical among all European countries and the degree of discrimination perceived could be different. Therefore, in order to see whether there are any potential opportunities to develop suitable policies and rules to address migration discrimination accordingly. The following hypothesis is made:

H3:There are distinct groups of migrants who share perception of similar discrimination based on the demographic background of migrants in those countries.

3 Data

3.1 Data pre-processing

This study will utilize the dataset from Eurostat, which was collected in the year 2021 from 29 countries including EU members and non-EU members through the European Union Labour Force Survey (EU LFS). The survey was conducted by randomly rotating samples from migrants aged between 15 and 74 about the type of discrimination they have perceived when seeking a job. The dataset gathered the number of migrants (measured in thousands of persons) who fall under different combinations of gender, age, country of birth, hosting country, educational attainment, and perception of types of discrimination experienced by the labour force. The dataset originally consisted of some observations with missing values in certain variables. As the sample size is relatively large, it has been determined that these observations will be excluded from the analysis, as missing values relate to individual demographic characteristics that cannot be inferred on the basis of available data since they related to personal circumstances. After removing the observations with missing values, there were 78534 thousand of observations and 27 countries remaining.

3.2 Response variable

The response variable is the type of discrimination that the respondent perceives. Although most response names are self-explanatory, some may require further clarification for better categorization. Discrimination type such as “No suitable job available” refers mainly to the mismatch between the skills of migrants and the availability of jobs. “Never sought work or never worked” refers to having no prior work experience. Table 1 shows the total observations for each type of perceived discrimination. 55.54 per cent of migrants reported that they have not experienced any discrimination, and 22.43 per cent reported they have experienced only one of the discrimination type and 20.03 per cent reported that they have experienced more than one discrimination type.

Response variable	**Total observations (thousands)**
Citizenship or residence permit	356.6
Discrimination due to foreign origin	500.1
Lack of language skills	2363.5
Lack of recognition of qualifications	1346.0
Language skills, qualifications, citizenship, foreign origin, job and other barriers	15959.9
Never sought work or never worked	9270.6
No suitable job available	1263.0
None	44268.0
Other	4370.2

Table 1: Response variable with associated number of observations (in thousands person)

3.3 Predictors

In this study, the predictors consist of the demographic factors of the migrants, each predictor being categorical with either nominal or ordinal in nature and consisting of different number of levels. The detailed predictors and associated levels are shown below:

Predictor	Levels
Age	15-24 years old, 25-54 years old, 55-74 years old
Gender	Females, Males
Educational attainment	Less than primary, primary and lower secondary education, Tertiary education, Upper secondary and post-secondary non-tertiary education
Country of birth	EU27 except reporting country, Foreign country, Non-EU27 countries (from 2020) nor reporting country
Hosting country	Austria, Belgium, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland

Table 2: Predictors and associated levels

4 Methodology

Since both the response variable and the predictors are categorical with unordered categories, this study will employ multinomial logistic regression and latent class analysis, which are the statistical methods designed for analyzing categorical data.

4.1 Multinominal Logistic Regression

Before starting the modelling process, the correlation between each pair of predictors is tested by Cramér's, which is a measurement designed particularly to examine the association between categorical variables. This helps in detecting potential collinearity and necessity of either removing highly associated variables or add interaction terms to explore the joint effect of variables.

The modelling process will start by selecting the appropriate predictors to be included in the model using fit statistics. This is done by fitting nested models (adding one variable at a time to the model) and measuring the reduction in the Akaike Information Criterion

(AIC) and Bayesian Information Criterion (BIC), which measures the balance between model fit against complexity. The process will then continue by testing the statistical significance of the predictors through evaluating the performance of the likelihood ratio test (measures whether the full model fits the data significantly better than the null model) and McFadden's R^2 (measures how much better the model is than the null model). Furthermore, the overall predictive ability of the model will also be evaluated using the confusion matrix (compare the model predictions with actual observations) to test whether the current model is able to predict the discrimination response well.

After fitting the model, predictors are analyzed to estimate their effect on discrimination responses. In particular, the effect of gender will be assessed to investigate whether female migrants are more likely to perceive discrimination and how it varies between different types through the fit statistics. The analysis will also examine the effect of educational attainment on the probability of perceiving discrimination to test whether there is a phenomenon of "integration paradox" in the European labour market.

4.2 Latent class analysis (LCA)

Latent Class Analysis is a model-based clustering method for categorical variables that fits probabilistic models to the data and uses multiple discrete observed variables as indicators to model discrete latent variables. This study employs LCA to identify unobserved subgroups of migrants based on their discrimination experience and demographic characteristics by defining all variables including *Discrimination perceived* as indicators of the model except *Hosting country.* In this section, the subset of dataset is used by removing the observations that have not reported any discrimination experienced to focus on analyzing the discrimination pattern.

Hosting country will be treated as a covariate to examine how different countries in which the migrants reside would influence their latent class membership with the goal of finding out if policy recommendations to address discrimination can be tailored to the unique context of each country. This method also helps preserve the valuable information provided by the dataset since in the dataset, certain countries have more observations compared to others, creating an imbalance that could dominate the analysis and resulted in biased latent class assignments. By incorporating *Hosting country* as a covariate, its effect is taken into account in the model without allowing the country sample size directly impact the latent class structure.

The analysis is carried out in four steps. First, all categorical variables undergo numerical encoding. The ordinal variables (*Age* and *Educational attainment*) are mapped to sequential integers preserving their inherent ordering, and the remaining nominal variables assigned arbitrary but consistent integer values. Secondly, before model fitting, models with 1 to 10 classes are run fit statistics are compared to decide the number of classes should be used in the model. The fit statistics include AIC and BIC, and entropy, which evaluates how well the latent classes are separated in the model. The entropy value usually ranges from 0 to 1 and a value greater than 0.8 indicates a good classification of individual observations into classes. After fitting the model with the optimal number of classes, the proportion that each class takes in each level of indicators will be extracted to first analyze the characteristics of each class and determine whether the latent classes have been well-distinguished. To further assess the effect of *Hosting country* in which the migrant resides on the likely latent class to which it will belong, the predicted probability of migrants from a given hosting country in each class will be computed using the Softmax function and assign the countries based on the highest probability:

$$P(C_k|H_p) = \frac{\exp^{\text{log-odds}_k}}{1 + \sum_{k=2}^{K} \exp^{\text{log-odds}_k}}$$

Where C_k represents the kth latent class and log-odds$_k$ represents the coefficient estimate of the multinomial logistic regression for each class and each hosting country.

5 Results

5.1 Test of collinearity

Below shows the result of Cramér's V regarding the correlation between each pair of predictors. The result shows that in general the predictors exhibit negligible or weak associations (between 0 and 0.2). This implies that the predictors are largely independent of each other. Potential associations are shown to be between *Hosting country* and *Age*,*Educational attainment* and *Hosting country*, *Country of birth* and *Hosting country.* Therefore, in order to account for these potential associations, two-way interaction terms are included in the regression model.

	Country of birth	Educational attainment	Age	Sex	Hosting country
Country of birth	1.0000	0.0698	0.0308	0.0104	0.1340
Educational_attainment	0.0698	1.0000	0.1354	0.0448	0.1932
Age	0.0308	0.1354	1.0000	0.0337	0.1604
Sex	0.0104	0.0448	0.0337	1.0000	0.0528
Hosting country	0.1340	0.1932	0.1604	0.0528	1.0000

Table 3: Cramér's V scores for each pair of predictors

5.2 Optimal model fitting and test of model performance

The AIC and BIC scores were calculated on each nested model and the result is shown in Figure 1. The plot shows that starting from the null model, both scores only decreased slightly after *Country of birth*, *Educational attainment* and *Age* were included and decreased significantly after adding *Hosting country* predictor. This indicates that it contributed substantially to the model. The consideration of potential interactions between *Educational attainment* with *Age* and *Hosting country* did not show any significant improvement in the model fit, as AIC remained relatively constant after adding these two interactions while BIC showed a significant increase. Therefore, the final model will only include all predictors given by the dataset without any interaction terms.

The likelihood ratio test was conducted to compare the fit of the chosen model with the null model. The result of the test revealed a statistically significant improvement in model fit, with a deviance difference of 30585.2765 and a p-value of 0. This indicated that all the predictors included have a statistically significant effect on the response at the 5 per cent significance level and have a meaningful impact on the response. However, McFadden's R^2 had a result of 0.1474, indicating that its effect size was modest. This was further supported by an accuracy rate of 58 per cent from Confusion Matrix, which indicated that the model has suboptimal predictive performance which could be improved by adding more key predictors. However, this does not undermine the value of inference that could be drawn from the model, the statistical significance of the predictors to the response still remains valid.

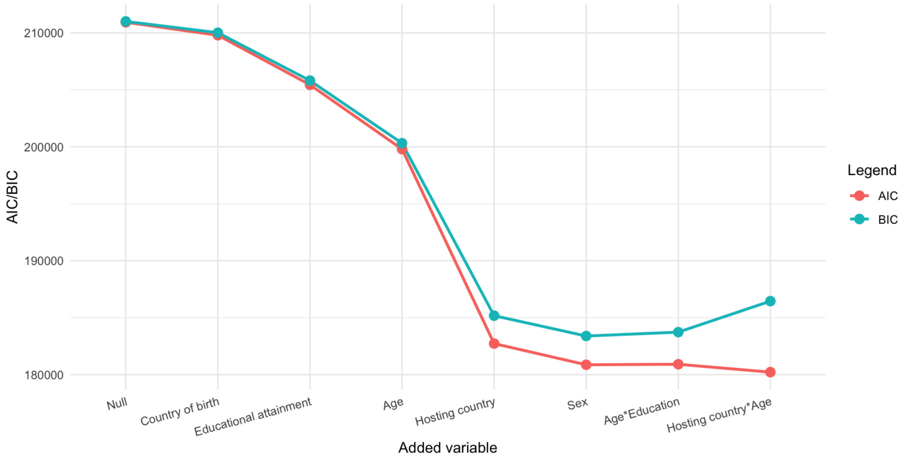

Figure 1: Change in AIC/BIC for nested model

5.3 Effect of gender on perceived discrimination

The effect of gender on perceived discrimination was extracted from the model results. The result showed that *Gender* has statistically significant effect on most types of discrimination (except *Citizenship or residence permit*) at the 5 per cent level, the standard errors measuring the variability of the coefficient estimates were small. For responses except *Citizenship or residence permit*, the coefficient estimates were positive, indicating that being a female migrant increases the odds of perceiving a particular type of discrimination (compared to reference level *None*). In particular, *Never sought work or never worked* showed the highest absolute value of the estimate, which revealed that the odds of perceiving *Never sought work or never worked* were 3.9440 times higher for female migrants compared to male migrants. Therefore, the result confirms the first hypothesis that female migrants are more likely to perceive discrimination compared to male migrants for reasons such as lack of qualifications, lack of prior job experience, or mismatch between skill and job requirements. This result also indirectly revealed that despite the increasing number of female migrants over time, they continue to face persistent and disproportionate discrimination, structural barriers to gender equality remain prevalent, limiting their access to opportunities and fair treatment.

y.level	term	estimate	std.error	statistic	p.value
Citizenship or residence permit	SexFemales	-0.0465	0.0945	-0.4918	0.6229
Discrimination due to foreign origin	SexFemales	0.2823	0.0862	3.2754	0.0011**
Lack of language skills	SexFemales	0.4628	0.0443	10.4568	0***
Lack of recognition of qualifications	SexFemales	0.7728	0.0602	12.8395	0***
Language skills, qualifications, citizenship, foreign origin, job and other barriers	SexFemales	0.3513	0.0195	18.0077	0***
Never sought work or never worked	SexFemales	1.3722	0.0284	48.2950	0***
No suitable job available	SexFemales	0.5974	0.0588	10.1552	0***
Other	SexFemales	0.2728	0.0340	8.0240	0***

[a] Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (Female as reference category)

Table 4: Coefficient estimates, standard error and p-value for Sex variable

5.4 Effect of educational attainment on perceived discrimination

Table 5 extracted the effect of educational attainment on response. P-value results showed that *Educational attainment* has statistically significant effect on most types of discrimination (except for *Citizenship or residence permit*) at the 0.1 per cent level. As we have observed from previous section, both predictors (*Educational attainment* and *Gender*) have not shown significant effect on *Citizenship or residence permit* response. The main reason was due to the small sample size and the model had less information to be able to estimate the relationship between various predictors and this type of response.

y.level	term	estimate	std.error	statistic	p.value	odds_ratio
Citizenship or residence permit	Tertiary education	-0.0381	0.1247	-0.3059	0.7597	0.9626167
Citizenship or residence permit	Upper secondary and post-secondary non-tertiary education	0.1517	0.1094	1.3859	0.1658	1.1638110
Discrimination due to foreign origin	Tertiary education	0.6332	0.1241	5.1028	0***	1.8836286
Discrimination due to foreign origin	Upper secondary and post-secondary non-tertiary education	0.7952	0.1124	7.0717	0***	2.2148839
Lack of language skills	Tertiary education	-0.3161	0.0529	-5.9748	0***	0.7289865
Lack of language skills	Upper secondary and post-secondary non-tertiary education	-0.6462	0.0556	-11.6277	0***	0.5240333
Lack of recognition of qualifications	Tertiary education	2.6923	0.1200	22.4332	0***	14.7655978
Lack of recognition of qualifications	Upper secondary and post-secondary non-tertiary education	1.4323	0.1261	11.3586	0***	4.1883213
Language skills, qualifications, citizenship, foreign origin, job and other barriers	Tertiary education	0.5289	0.0252	20.9562	0***	1.6970645
Language skills, qualifications, citizenship, foreign origin, job and other barriers	Upper secondary and post-secondary non-tertiary education	0.1672	0.0246	6.8011	0***	1.1819906
Never sought work or never worked	Tertiary education	-1.0295	0.0362	-28.4314	0***	0.3571855
Never sought work or never worked	Upper secondary and post-secondary non-tertiary education	-0.9199	0.0304	-30.2648	0***	0.3985589
No suitable job available	Tertiary education	1.2879	0.0815	15.8054	0***	3.6251657
No suitable job available	Upper secondary and post-secondary non-tertiary education	0.6371	0.0812	7.8439	0***	1.8909891
Other	Tertiary education	0.5072	0.0458	11.0771	0***	1.6606349
Other	Upper secondary and post-secondary non-tertiary education	0.2470	0.0398	6.2025	0***	1.2801791

[a] Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (Less than primary, primary and lower secondary education as reference category)

Table 5: Coefficient estimates, standard error and p-value for Educational attainment variable

For both levels of educational attainment (*Upper secondary and post-secondary non-tertiary education* and *Tertiary education*) relative to *less than primary, primary and lower secondary education*, the coefficient estimates were positive for majority of the responses. This revealed that educated migrants were more likely to report having experienced discrimination. More importantly, the coefficient estimates are generally greater for migrants with tertiary education compared to those with Upper secondary and post-secondary non-tertiary education and the odds of highly educated migrants (with tertiary education) reporting multiple discrimination experienced were 1.6971 times higher than those less educated whioch supports the "integration paradox" argument by Verkuyten M.

5.5 Optimal number of classes for LCA

The fit statistics were obtained after fitting the models with number of classes from 1 to 10. The plot revealed that both AIC and BIC decreased as the number of classes increased from 1 to 3, indicating a model fit improvement. However, between 3 and 4 classes, AIC showed a slight decrease while BIC showed a slight increase, and both scores started to fluctuate after 4 classes. By considering only both AIC, BIC and entropy scores, the optimal number of classes should be between 3 and 4. The plot also included the entropy score for each number of classes. The entropy score was higher with 3 classes and the score exceeds to threshold of 0.8, indicating a better classification of observations relative to when the number of classes was 4. Therefore, the optimal number of classes chosen for the latent class analysis should be 3.

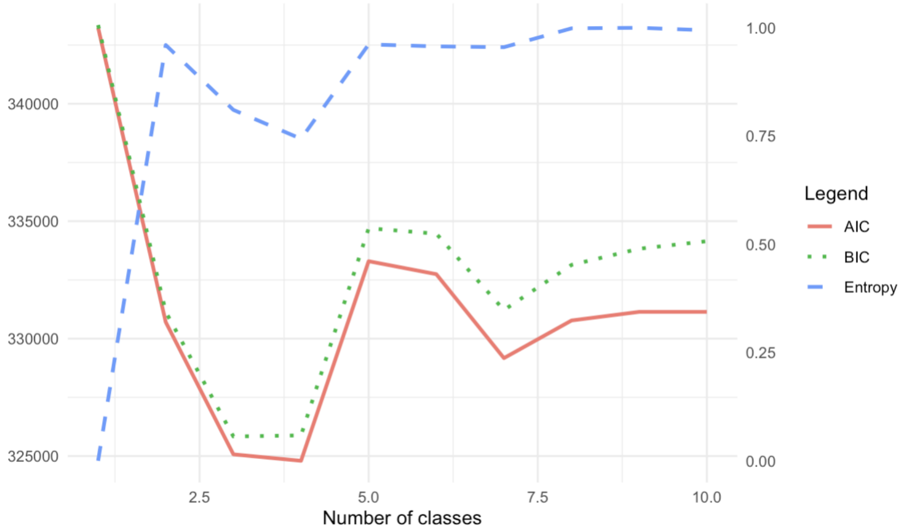

Figure 2: AIC,BIC and Entropy scores for Latent Class Analysis with different number of classes

5.6 Extracting characteristics of each latent class

Figure 3 shows the proportion each latent class takes in each level of indicators. From the plots it can be observed that the proportions each class took for different countries of birth were similar across all classes with around 60 per cent of the migrants being from foreign countries outside Europe. This implied that *Country of birth* was not a dominant factor that could distinguish latent classes. In terms of *Age*, Class 1 and 2 were characterized by migrants mostly between 25-54 years old (75-80 percent). Class 3 showed a more distinguishable pattern consisting of relatively equal proportions of all age groups.

In terms of discrimination patterns, *Citizenship or residence permit* and *Discrimination due to foreign origin* were not observed between all classes. This was due to small observations of these two responses in the dataset. All classes had a high proportion of migrants perceiving multiple discrimination relative to other types, except Class 3 which consisted of only *Never sought worked or never worked*. There were also other types of discrimination detected in the classes, with Class 1 having 30 per cent of migrants perceiving *Other* types of discrimination and Class 2 had a higher proportion for *Lack of language skills* among three classes. All classes showed different patterns of *Educational attainment*.

When comparing Class 1 and Class 2, both classes had a relatively equal proportion of migrants who received either less than primary or secondary education but Class 1 had a higher proportion of migrants compared to Class 2. Class 3 consisted mainly of migrants with less than primary and secondary education compared to the other classes. In terms of gender distribution, Class 1 and 2 shared equal gender proportions, Class 3 was female dominated. Hence, the features of each class can be summarized as following:

- Class 1: Relatively less educated migrants in the working population who have experienced multiple discrimination and some other types not mentioned (which could be race-related).

- Class 2: Educated migrants in the working population with large proportion have reported experiencing more than one type of discrimination and minor proportions reported lacking of language skills or qualifications.

- Class 3: Less educated and female-dominated migrants among all different age groups who could not find a suitable job due to lack of job experience.

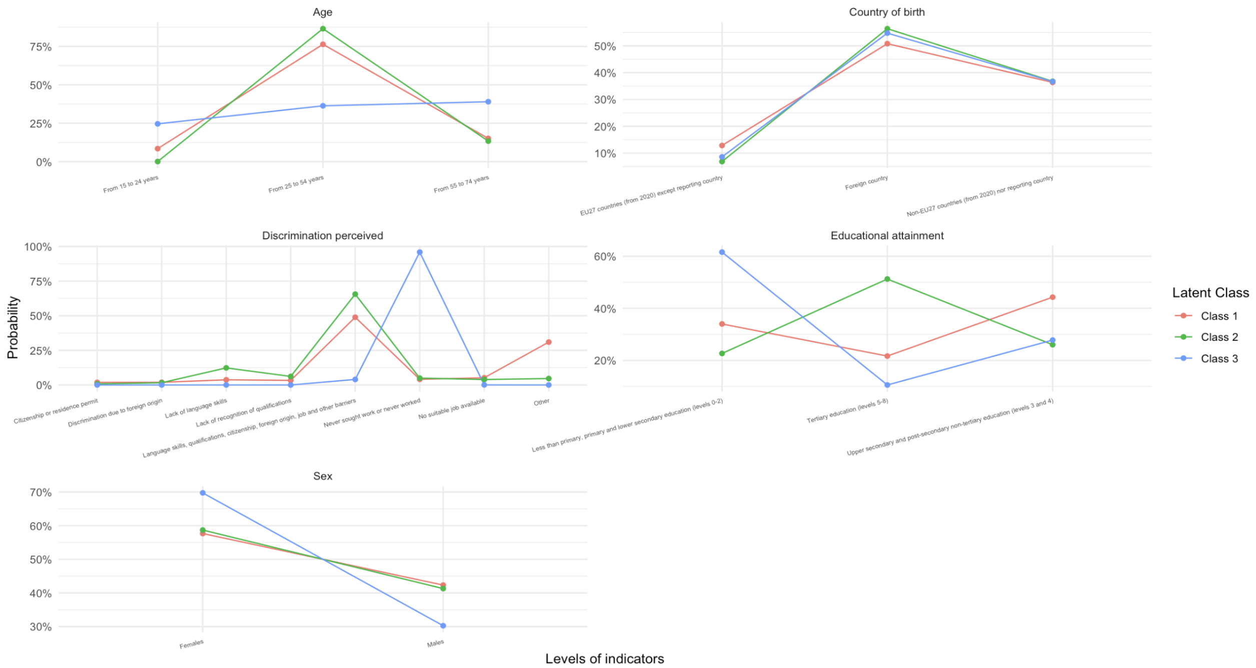

Figure 3: Latent Class Response Probabilities

After understanding the uniqueness of each latent class, the multinomial logistic coefficient estimates on the covariate *Hosting country* were extracted. The probability of observations in a given hosting country under each class provides a relative insight on which class that migants in each hosting country would mostly likely belong to.

Class	Countries
1	Denmark, Estonia, Finland, France, Greece, Ireland, Netherlands, Norway, Spain, Sweden, Switzerland
2	Lithuania, Luxembourg, Portugal
3	Belgium, Croatia, Cyprus, Czechia, Germany, Hungary, Malta, Slovenia

Table 6: Country assignment to latent classes based on coefficient estimates

6 Policy recommendation and conclusion

Following the result of the Latent Class Analysis, three classes were observed together with the most likely class in which migrants from each hosting country are likely to be. This result allows policies to be designed to address the discrimination situation in each latent class. Overall, for the countries in the first two classes, a large proportion of migrants have reported that they experienced multiple discrimination, this implies that strong anti-discrimination laws should be enhanced to provide migrants with more job opportunities and stronger workplace protections. More specifically, for the first latent class, the source of discrimination comes mainly from the lack of education of migrants based on the general demographic characteristics of the class. Lack of education brings local language and job mismatching challenges; therefore, approaches to help address this would be to provide subsidized skills and language training courses or within-workplace training programmes tailored to the needs of labour market. For the second class, migrants are mostly educated but still experience a lack of language skills and qualifications. This could be addressed by providing training courses as mentioned previously while developing a more simplified skill recognition procedure to mutually recognize the qualifications that migrants obtained from their country of birth. Class 3 relates to the obstacles faced by uneducated female migrants when seeking for a job. This could be addressed from two perspectives. Firstly, more affordable daycare should be offered to female migrants who need to fulfill their family responsibilities to provide them with sufficient time for skill development and job seeking. Secondly, encourage female migrant-owned businesses to reduce the potential discrimination perceived from traditional employment channels.

Taken together, this study discovers how the demographic background of migrants would potentially affect the likelihood of perceiving discrimination in particular the hosting country in which the migrant resides. As we observed from the result of multinomial logistic regression, the demographic factors have not fully explained the response pattern of discrimination, other potentially underlying reasonse should be further investigated and studied to understand if there are any unique causes in each country leading to such discrimination perception and more tailored policies can be developed further to help address the situation in individual country level. As part of the labour force participants these should be addressed in a country-level to minimize the probability of discouraging the migrants contributing to the labour market.

7 R codes

```r
#Response variable
suppressMessages(library(dplyr))
suppressMessages(library(knitr))
suppressMessages(library(kableExtra))
library(xtable)
#Import the dataset
dataset <- read.csv("/Users/xiaojinghuang/Desktop/STA5071Z/
estat_lfso_21obst01_filtered_en-2 copy.csv")
dataset <- dataset[,c(4:8,10,12)]
dataset <-dataset[complete.cases(dataset),]
#Removing missing observations
dataset <- na.omit(dataset)
colnames(dataset) <- c("Country_of_birth","Educational_attainment",
"Discrimination_perceived",
"Age","Sex","Hosting_country","Observed_population")
dataset <- dataset %>%
  mutate(across(-ncol(dataset), as.factor))

responses <- data.frame(
  Response_variable = dataset$Discrimination_perceived,
  Total_observations_in_thousands = dataset$Observed_population)

responses_table <- responses %>%
  group_by(Response_variable) %>%
  summarize(Total_observations_in_thousands =
  sum(Total_observations_in_thousands))

knitr::kable(responses_table, format = "latex",
booktabs = TRUE,caption = "Response variable with associated
number of observations (in thousands person)") %>%
    kableExtra::kable_styling(latex_options =
    c("hold_position", "scale_down"), font_size = 10) %>%
    kableExtra::column_spec(2, width = "6cm", latex_valign = "m")

#Predictors
Age <- factor(c("15-24 years old", "25-54 years old",
"55-74 years old"))
Gender <- factor(c("Male", "Female"))
Educational_attainment <- factor(c("Less than primary,
primary and lower secondary education (Level 0-2)",
"Upper secondary and post-secondary non-tertiary education",
"Tertiary education (Level 5-8)"))
Country_of_birth <- factor(c("EU27 except reporting country",
"Non-EU27 countries (from 2020) nor reporting country",
"Foreign country"))
```

```
Hosting_country <- factor(c("Spain","Italy","Hungary","Netherlands",
"Austria","Belgium","Switzerland","Greece","Croatia","France",
"Luxembourg","Cyprus","Czechia","Germany","Portugal","Finland",
"Estonia","Norway","Slovenia","Sweden","Denmark","Slovakia",
"Ireland","Lithuania","Malta","Poland","Latvia"))

predictors <- data.frame(
  Predictor = c("Age", "Gender","Educational_attainment",
  "Country_of_birth",
  "Hosting_country"),
  Levels = c(paste(levels(Age), collapse = ", "),
  paste(levels(Gender), collapse = ", "),
  paste(levels(Educational_attainment), collapse = ", "),
  paste(levels(Country_of_birth), collapse = ", "),
  paste(levels(Hosting_country), collapse = ", ")))

knitr::kable(predictors, format = "latex", booktabs = TRUE,
caption = "Predictor levels") %>%
  kableExtra::kable_styling(latex_options =
  c("hold_position", "scale_down"), font_size = 10) %>%
  kableExtra::column_spec(2, width = "12cm", latex_valign = "m")

#Test of collinearity
suppressMessages(library(vcd))
#Expand the dataset and remove the response variable
dataset_exp <- dataset[rep(1:nrow(dataset),
dataset$Observed_population),-ncol(dataset)]
dataset2 <- dataset_exp[,c(1:2,4:6)]
cramers_v_matrix <- function(data) {
  n <- ncol(data)
  cramers_matrix <- matrix(NA, nrow = n, ncol = n)
  colnames(cramers_matrix) <- colnames(data)
  rownames(cramers_matrix) <- colnames(data)
  for (i in 1:n) {
    for (j in 1:n) {
      if (i == j) {
        cramers_matrix[i,j] <- 1  # Diag=1
      } else {
        a <- table(data[,i], data[,j])
        cramers_matrix[i,j] <- assocstats(a)$cramer
      }
    }
  }
  return(cramers_matrix)
}
# Calculate Cramer's V for all pairs
cramers <- as.data.frame(cramers_v_matrix(dataset2))
cramers <- round(cramers,4)
knitr::kable(cramers, format = "latex", booktabs = TRUE,
             caption = "Cramer's V scores for each pair of predictors") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"),
                            font_size = 10) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")

#Optimal model fitting and model performance test
suppressWarnings({
suppressMessages(library(nnet))
require(broom, quietly = TRUE)
suppressMessages(library(ggplot2))
suppressMessages(library(tidyr))
require(pscl, quietly = TRUE)
set.seed(123)
#Model fitting by adding one variable at a time
null_model <- multinom(Discrimination_perceived ~ 1,
                       data=dataset_exp,
                       trace=FALSE)
model_int1 <- multinom(Discrimination_perceived ~Country_of_birth,
                      data=dataset_exp,
                       trace=FALSE)
model_int2 <- multinom(Discrimination_perceived ~ Country_of_birth +
                         Educational_attainment,
                       data=dataset_exp,
                       trace=FALSE)
model_int3 <- multinom(Discrimination_perceived ~ Country_of_birth +
                        Educational_attainment + Age,
                       data=dataset_exp,
                       trace=FALSE)
model_int4 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age + Hosting_country,
                       data=dataset_exp,trace=FALSE)
model_int5 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age + Hosting_country +
                         Sex, data=dataset_exp,trace=FALSE)
model_int6 <- multinom(Discrimination_perceived ~ Country_of_birth+
                         Educational_attainment+ Age  + Hosting_country + Sex+
                       Educational_attainment*Age,
                       data=dataset_exp,nnet.MaxNWts =6000,trace=FALSE)
model_full <- multinom(Discrimination_perceived ~ Country_of_birth +
                         Educational_attainment+ Age  + Hosting_country + Sex +
                         Educational_attainment*Age +
                         Hosting_country*Educational_attainment,
                       data=dataset_exp,nnet.MaxNWts =6000,trace=FALSE)

#BIC calculation
```

```
null <- BIC(null_model)
first <- BIC(model_int1)
second <- BIC(model_int2)
third <- BIC(model_int3)
fourth <- BIC(model_int4)
fifth <- BIC(model_int5)
sixth <- BIC(model_int6)
full <-  BIC(model_full)
#AIC calculation
nulla <- AIC(null_model)
firsta <- AIC(model_int1)
seconda <- AIC(model_int2)
thirda <- AIC(model_int3)
fourtha <- AIC(model_int4)
fiftha <- AIC(model_int5)
sixtha <- AIC(model_int6)
fulla <-  AIC(model_full)
AIC <- c(nulla,firsta,seconda,thirda,fourtha,fiftha,sixtha,fulla)
BIC <- c(null,first,second,third,fourth,fifth,sixth,full)
added_variables <- c("Null","Country of birth","Educational attainment",
"Age", "Hosting country", "Sex","Age*Education","Hosting country*Age")
added_variables <- factor(added_variables)
table <- data.frame(added_variables, AIC,BIC)
table$added_variables <- factor(table$added_variables,
                                levels = table$added_variables)

#Plot
table_long <- pivot_longer(table, cols = c(AIC, BIC), names_to = "Line",
                           values_to = "Value")
                           #Change the struture of data
ggplot(table_long, aes(x = factor(added_variables),
y = Value, group = Line,color = Line)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(caption = "Figure 1: Change in AIC/BIC for nested model",
       x = "Added variable",
       y = "AIC/BIC",
       color = "Legend") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1,vjust = 1),
    plot.caption = element_text(hjust = 0.5))

# Final model
dataset_exp$Discrimination_perceived <-
factor(dataset_exp$Discrimination_perceived, ordered = FALSE)
dataset_exp$Discrimination_perceived <-
relevel(dataset_exp$Discrimination_perceived, ref="None")
```

```
dataset_exp$Sex <- factor(dataset_exp$Sex, ordered = FALSE)
dataset_exp$Sex <- relevel(dataset_exp$Sex, ref="Males")
#Change reference level

model_final <- multinom(Discrimination_perceived ~ Country_of_birth +
                    Educational_attainment + Age + Sex + Hosting_country,
                       data=dataset_exp,trace=FALSE)
library(caret)
library(pscl)
require(caret, quietly = TRUE)

# Confusion matrix
pred <- predict(model_final, type = "class")
actual_dis <-  dataset_exp$Discrimination_perceived
confusion <- confusionMatrix(pred, actual_dis)

# Likelihood ratio test
null_deviance <- deviance(null_model)
full_deviance <- deviance(model_final)
test <- null_deviance - full_deviance
diff <- length(coef(model_final)) - length(coef(null_model))
p_value <- 1 - pchisq(test, diff)

# Mcfadden R^2
mcfadden <- 1 - logLik(model_final)/logLik(null_model)
})

#Effect of gender on perceived discrimination
suppressMessages(library(nnet))
require(broom, quietly = TRUE)
suppressMessages(library(ggplot2))
#Fit the model
colnames(tidy_results)
tidy_results <- tidy_results %>% mutate(
    estimate = round(estimate, 4),  ##Round to 4 dec.
    std.error = round(std.error, 4),
    statistic = round(as.numeric(statistic), 4),
    p.value = round(p.value, 4),
    p.value = paste0(p.value,case_when(
                                  p.value < 0.001 ~ "***",
                                  p.value < 0.01 ~ "**",
                                  p.value < 0.05 ~ "*", #Scientific not
                                  TRUE ~ " "
                                )))
tidy_results <- tidy_results %>% mutate(
   statistic = as.numeric(formatC(statistic,
   format = "f", digits = 4)))
```

```
table3 <- knitr::kable(tidy_results, format = "latex", booktabs = TRUE,
             caption = "Coefficient estimates,
             standard error and p-value for all he predictors") %>%
  kableExtra::kable_styling(latex_options = c("hold_position",
  "scale_down"),
                            font_size = 12) %>%
  add_footnote("Note: *** p < 0.001, ** p < 0.01, * p < 0.05") %>%
  kableExtra::column_spec(2, width = "5cm", latex_valign = "m")
cat(table3)
#Hypothesis 1
H1_Sex <- tidy_results %>%
  filter(term == "SexFemales") #Extract results relating to gender
knitr::kable(H1_Sex, format = "latex", booktabs = TRUE,
             caption = "Coefficient estimates, standard error
             and p-value for Sex variable") %>%
  kableExtra::kable_styling(latex_options =
  c("hold_position", "scale_down"),
                            font_size = 12) %>%
  add_footnote("Note: *** p < 0.001, ** p < 0.01, * p < 0.05
  (Female as reference category)") %>%
  kableExtra::column_spec(2, width = "5cm", latex_valign = "m")

## Effect of educational attainment on perceived discrimination
# the effect of education predictor
H2_Education1 <- tidy_results %>%
  filter(grepl("Educational_attainment", term))
H2_Education1 <- H2_Education1 %>%
  mutate(odds_ratio = exp(estimate))
knitr::kable(H2_Education1, format = "latex", booktabs = TRUE,
             caption = "Coefficient estimates, standard error
             and p-value for
             Educational attainment variable") %>%
  kableExtra::kable_styling(latex_options =
  c("hold_position", "scale_down"),font_size = 12) %>%
  add_footnote("Note: *** p < 0.001, ** p < 0.01, * p < 0.05
  (Less than primary, primary and lower secondary education
  as reference category)") %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")

## Optimal number of classes for LCA
# Transform the dataset
suppressMessages(library(poLCA))
# Dummy code the dataset
dataset_exp <- dataset[rep(1:nrow(dataset),dataset$Observed_population),
-ncol(dataset)]
filter_data <- dataset_exp %>%
  filter(Discrimination_perceived != "None")
```

```r
filter_data$Age <- factor(
  filter_data$Age ,
  levels = c("From 15 to 24 years", "From 25 to 54 years",
  "From 55 to 74 years"),  # Explicit order
  ordered = TRUE
)
filter_data$Educational_attainment <- factor(
  filter_data$Educational_attainment,
  levels = c("Less than primary, primary and lower secondary
  education (levels 0-2)", "Upper secondary and post-secondary
  non-tertiary education (levels 3 and 4)", "Tertiary education (levels 5-8)"),
  ordered = TRUE
)
filter_data$Discrimination_perceived <-
as.factor(filter_data$Discrimination_perceived)


ggplot(filter_data, aes(x = Hosting_country,
fill = Discrimination_perceived)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Hosting Country", y = "Count of Discrimination Types",
       title = "Discrimination Types by Hosting Country") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size = 5))
#Filtered the data results in levels of indications being reduced
filter_data$Discrimination_perceived <-
droplevels(filter_data$Discrimination_perceived)
filter_data$Hosting_country <- droplevels(filter_data$Hosting_country)
#Run the model for different number of classes and extract AIC,BIC and entropy
AIC_l2 <- c()
BIC_l2 <- c()
entropy2 <- c()
for (k in 1:10) {
  set.seed(123)
  model2 <- poLCA(cbind(Country_of_birth,Educational_attainment,
  Discrimination_perceived,Age,Sex)~Hosting_country, data=filter_data,
  nclass=k,maxiter = 5000,nrep=5)
  AIC_l2[k] <- model2$aic
  BIC_l2[k] <- model2$bic
  entropy1 <- function(p) {       #entropy calculation
  -sum(p * log(p), na.rm = TRUE)
}
  avg_entropy<- mean(apply(model2$posterior, 1, entropy1), na.rm = TRUE)
  entropy2[k] <- 1 - (avg_entropy / log(k))
}
tablev2 <- data.frame(1:10,AIC_l2,BIC_l2,entropy2)
```

```
tablev2$entropy2[1] <- 0
tablev2$e_rescale <- (tablev2$entropy2 - min(tablev2$entropy2))
/ (max(tablev2$entropy2)
                    - min(tablev2$entropy2)) * (max(tablev2$AIC_l2)
                    - min(tablev2$AIC_l2)) + min(tablev2$AIC_l2)

# Plot with dual y-axis with AIC, BIC and Entropy
ggplot(tablev2, aes(x = X1.10)) +
  geom_line(aes(y = AIC_l2, color = "AIC"), size = 1) +
  geom_line(aes(y = BIC_l2, color = "BIC"), size = 1,
  linetype = "dotted") +
  geom_line(aes(y = e_rescale, color = "Entropy"), size = 1,
  linetype = "dashed") +
  scale_y_continuous(name = NULL,
    sec.axis = sec_axis(
      ~ (. - min(tablev2$AIC_l2)) * (max(tablev2$entropy2) -
      min(tablev2$entropy2)) / (max(tablev2$AIC_l2) - min(tablev2$AIC_l2)) + min(tabl
      name = NULL
    )
  ) +
  labs(x = "Number of classes",caption="Figure 2: AIC,BIC and
  Entropy scores for Latent Class Analysis with different number of classes",
  color = "Legend") +
  theme_minimal() +
  theme(
    plot.caption = element_text(hjust = 0.5)
  )

## Extracting the charcteristics of the latent classes

# Fit the model-hosting country as covariate
modelv2 <- poLCA(cbind(Country_of_birth,Educational_attainment,
Discrimination_perceived,Age,Sex)~Hosting_country, data=filter_data,
nclass=3,maxiter = 6000,nrep=10)

# Plot the distribution of indicator levels for each class
probs2 <- modelv2$probs
predicted_class <- modelv2$predclass
response2 <- list()
for(v in names(probs2)){
  prob2 <-t(probs2[[v]])
  dataf2 <- as.data.frame(prob2)
  colnames(dataf2) <- paste0("Class ", seq_len(ncol(dataf2)))
  dataf2$Level <- rownames(dataf2)
  dataf2$Indicator <-v
  response2[[v]] <- dataf2
}
response_data2 <- bind_rows(response2)
long_data2 <- response_data2 %>%
  pivot_longer(cols = starts_with("Class "), names_to = "Latent_Class",
  values_to = "Probability")

facetlabs <- c("Age" = "Age", "Country_of_birth" =
"Country of birth", "Discrimination_perceived" =
"Discrimination perceived","Educational_attainment"=
"Educational attainment", "Sex"="Sex")
ggplot(long_data2, aes(x =Level, y = Probability,
group = Latent_Class, color = Latent_Class)) +
  geom_line() +
  geom_point() +
  facet_wrap(~Indicator, scales = "free",labeller =
  labeller(Indicator = facetlabs),nrow = 3,ncol = 2) +
  labs(x = "Levels of indicators" ,y = "Proportion", caption =
  "Figure 3: Latent Class Response Probabilities",color = "Latent Class") +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent) +
  theme(axis.text.x = element_text(angle = 15, hjust = 1,size = 5),
    plot.caption = element_text(hjust = 0.5))

#The effect of Hosting country covariate on assignment of observations
#in each latent class
#Coefficients of covariates
coeff <- modelv2$coeff
row <- rownames(coeff)
ncol(coeff)
coeffs <- data.frame(
  Class =rep(paste("Class ", 2:3), each=length(row)),
  Term = rep(row,times=2),
  Coefficient = as.vector(coeff))
knitr::kable(coeffs, format = "latex", booktabs = TRUE,
             caption = "The effect of Hosting country covariate
             on assignment of observations in each latent class") %>%
  kableExtra::kable_styling(latex_options =
  c("hold_position", "scale_down"),
                              font_size = 10) %>%
  add_footnote("Note:Class 1 and Hosting country Austria are the
  reference levels") %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")

# calculate the odds
odds <- exp(coeff)
reference_class <- rep(1,nrow(coeff))
odds <- cbind(reference_class,odds)
```

```r
colnames(odds) <- c("Class 1","Class 2","Class 3")
total <- rowSums(odds)
probability <- round(odds/total,4)
class <- apply(probability, 1, which.max)

# Design a table
country <- list(
  c("Denmark","Estonia","Finland","France","Greece","Ireland",
  "Netherlands","Norway","Spain","Sweden","Switzerland"),
  c("Lithuania","Luxembourg","Portugal"),
  c("Belgium","Croatia","Cyprus","Czechia","Germany","Hungary",
  "Malta","Slovenia")
)
country <- sapply(country, function(x) paste(x, collapse = ", "))

assign <- data.frame(
  Class = 1:3,
  Countries = I(country)
)

knitr::kable(assign, format = "latex", booktabs = TRUE,
             caption = "Country assignment to latent classes
             based on coefficient estimates") %>%
  kableExtra::kable_styling(latex_options = c("hold_position", "scale_down"),
                            font_size = 10) %>%
  kableExtra::column_spec(2, width = "6cm", latex_valign = "m")
```

References

Cortinovis et al., 2020.*Gendered migrant integration policies in the EU: Are we moving towards delivery of equality, non-discrimination and inclusion?* Available: https://cdn.ceps.eu/wp-content/uploads/2023/03/ITFLOWS$_{R}eport - on - gendered - migrant - integration - and - outcomes.pdf[10March, 2025]$.

Das S., Kotikula A., 2019. *Gender-based employment segregation: understanding causes and policy interventions.* Available: https://documents1.worldbank.org/curated/en/483621554129720460/pdf/Gender-Based-Employment-Segregation-Understanding-Causes-and-Policy-Interventions.pdf [8 March, 2025].

European Commission, 2025. *EU Immigration Portal.* Available: https://immigration-portal.ec.europa.eu/general-information/who-does-what$_{e}n[10March, 2025]$.

Giang H. Rima T., 2018. *The Labour Market Integration in Europe: New Evidence from Micro Data.* Available:

https://www.imf.org/en/Publications/WP/Issues/2018/11/01/The-Labor-Market-Integration-of-Migrants-in-Europe-New-Evidence-from-Micro-Data-46296 [10 March, 2025].

International Organization of Migration, 2020. *Chapter 3: Migration and Migrants: Regional Dimensions and Developments.* Available: https://worldmigrationreport.iom.int/what-wedo/world-migration-report-2024-chapter-3/europe [10 March, 2025].

Raijman R. Semyonov M., 1997. *Gender, ethnicity, and immigration: double disadvantage and triple disadvantage among recent immigrant women in the Israeli labor market.* Available: https://www.jstor.org/stable/190228 [12 March, 2025].

Sprengholz M.Schieckoff B.,2021.*The labor market integration of immigrant women in Europe: context, theory, and evidence.* Available: https://link.springer.com/article/10.1007/s43545-021-00279-3 [12 March, 2025].

Schaeffer M.Kas J., 2023.*The Integration Paradox: A Review and Meta-Analysis of the Complex Relationship Between Integration and Reports of Discrimination.* Available:https://journals.sagepub.com/doi/abs/10.1177/01979183231170809 [12 March, 2025].

United Nations, 2024. *International Migration Stock 2024.* Available: https://www.un.org/development/desa/pd/content/international-migrant-stock [12 March, 2025].

Verkuyten M., 2016. *The Integration Paradox: Empiric Evidence From the Netherlands.* Available: https://pubmed.ncbi.nlm.nih.gov/27152028/ [12 March, 2025]