German Criminal Statistics by Nationality

Data Management with R Final Project

Keri Hartman

15 December 2017

Introduction

Few topics generate so much heated debate, both in the political arena and around one's local *Stammtisch*, as the link between migration, and particularly the presence of asylum seekers, and crime. In my master's thesis, I plan to use panel data from the German Police Criminal Statistics (*Polizeiliche Kriminalstatistik*, hereafter: PKS) to explore the factors causing variation in crime rates among nationality groups across time. In addition to control variables, such as age, gender, family and educational structure, I plan to analyze the effect of such factors as immigration status (proportion of nationals considered asylum seekers, recognized refugees, with suspended deportation orders, etc. in Germany), religion, and geography in determining crime rates. I will also examine potential changes in the effects of these variables over time, specifically as a result of the more "uncontrolled" arrival of asylum seekers in 2015. This paper represents a first draft of this analysis. In it, I download and clean the PKS data as well as three other datasets with selected control variables, conduct exploratory analyses using a variety of plots and graphs, and run some initial regressions. Throughout, I not only describe my process but also discuss the limitations of my analyses in their current form, highlighting avenues for further refinement in the coming months.

Data Sources

Crime data

The main data source for this project were PKS statistics on arrested criminal suspects. These were available for the years 2012-2016 in Excel form from the **website** of the German Federal Criminal Police Office *(Bundeskriminalamt)*. Each year's statistics were contained in a separate Excel file; thus, I had five separate datasets after initial imputation, which I combined into a list before conducting further data cleaning. This allowed me to loop through the five annual datasets when running all subsequent data cleaning steps.

To get the datasets into tidy format, I reshaped them so that nationality groups (my panel units) formed the rows and crimes (my variables) formed the columns. I then converted the nationality group names into English. Finally, I combined the five datasets into one panel dataset containing all crimes committed by all nationality groups over the past five years.

In doing so, a few problems emerged. First, the crime descriptions used in the PKS have gone through a number of formatting changes over the past five years, and thus were of limited use for identifying specific crimes over time. Luckily, the PKS also includes codes for each crime that have not changed over the past five years; these were used to generate identifiers that were comparable across time. Second, while most crimes listed in each year's dataset can be assumed to have been committed in the same year, a few were "cold cases" (*Altfälle*) committed by citizens of countries

that no longer exist, such as the Sowiet Union or Yugoslavia. These nationality groups were omitted from the dataset for easier comparability.

Register data

Unfortunately, the PKS does not contain information on the age, gender, marital status, or immigration status of criminal suspects by nationality - despite the fact that these are important predictors of criminal behavior in general and are likely to differ across nationality groups. I thus was forced to rely on register data to obtain the values of these control variables for the general population of each nationality group residing in Germany. These were available in Excel form from the **website** of the German Federal Statistical Office (Destatis) for non-German citizens only. Comparable data for German citizens was not available.

Due to a paywall, only datasets on gender and marital status could be downloaded. These were again combined into a list before data cleaning. The two Destatis datasets were already in proper format in terms of rows and columns; thus, I only had to convert the header columns for each year into a proper year variable. Afterwards, I again converted the nationality group names into English, removed data for countries that no longer exist, and combined the data for countries that appeared multiple times (e.g. British Overseas Territories became part of an overall figure for the United Kingdom). In addition, the Excel file for marital status listed "-" instead of zero when no members of a nationality group had a given status, which was read as NA during imputation; these were converted to zeros to avoid data loss.

After cleaning the population data in this way, I combined the two control datasets with the main dataset. At this point, a conceptual problem arose due to the fact that the crime data represent a cumulative total of acts that occur throughout the year, whereas the population data are a fixed snapshot on 31 December of each year. I decided to match the number of crimes committed in each year with the population estimates for 31 December of that same year because I assumed that the population was more likely to grow than decline over the course of each year. Thus, taking end-of-year population figures should make my crime rate estimates more conservative. However, this remains just an assumption that will need to be tested during further work on this dataset for my master's thesis. A better method, not implemented at present, might be to take the mean of the population data for the beginning and end of a given year (i.e., 31 December 2011 and 31 December 2012 for 2012).

Geographical Data

Finally, in order to determine whether certain world regions tend to "specialize" in certain crimes, I added a variable classifying each country into the world region to which it belongs. This data was obtained from the a **nature conservation website** via web scraping because no suitable pre-existing database could be found. After importing the data and separating the imported strings into country and region variables, I converted the country names into the standard format used in the main dataset and merged the two datasets along this variable.

Exploratory Analyses

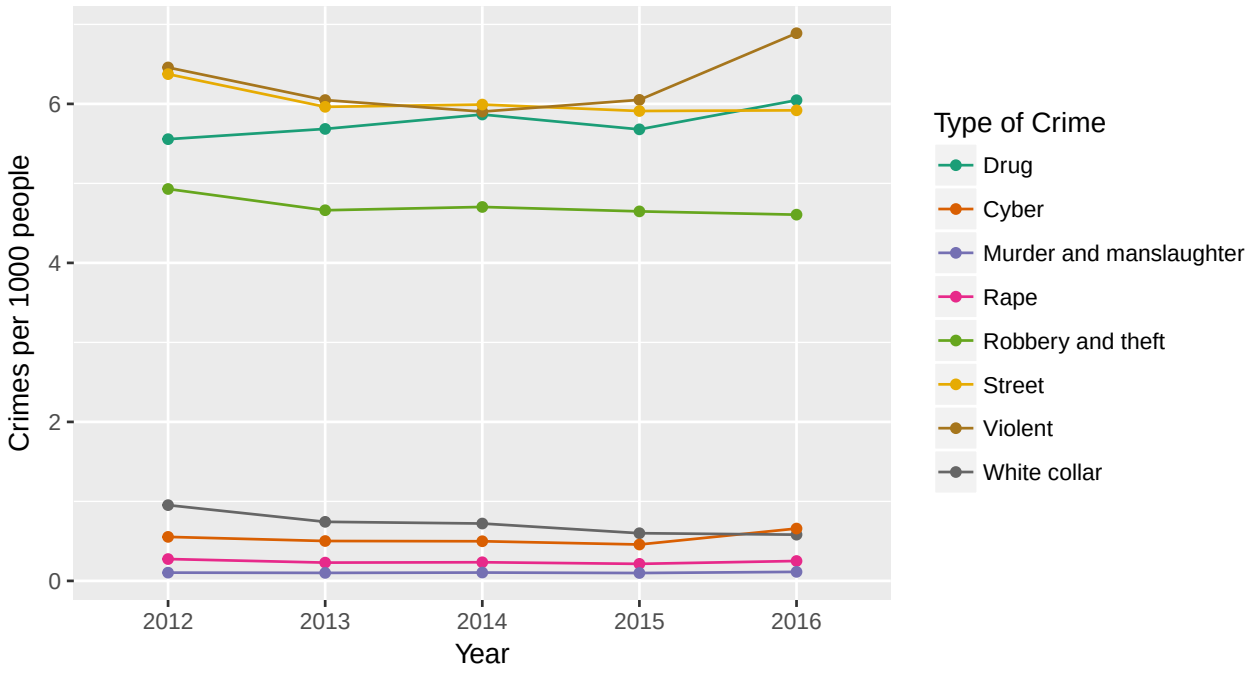
After finalizing my dataset, I conducted some further data cleaning and exploratory analyses. The PKS dataset contains information on 1109 separate crime codes, although some of these are summary codes collapsing different types of related crimes into nested hierarchies. In order to reduce this to a more manageable number, I created new variables for major crime categories of interest. Some of these already had PKS summmary codes (e.g., drug crimes, white collar or economic crimes), while others were my own creation (e.g., robbery and theft as a subset of street crimes). In addition, all crime variables were standardized to crime rates per 1000 people for easier interpretation - i.e., a value of five indicates that a given crime was charged five times per 1000 members of a given nationality group in a given year.

Importantly, the PKS datset records numbers of crimes, not numbers of unique individuals charged with crimes, meaning that a few especially prolific criminals can artificially inflate a given nationality's crime rate, especially when that nationality has a small overall population. For this reason, I excluded all nationality groups with a population of less than 20 from further analyses. This affected the following nationalities:

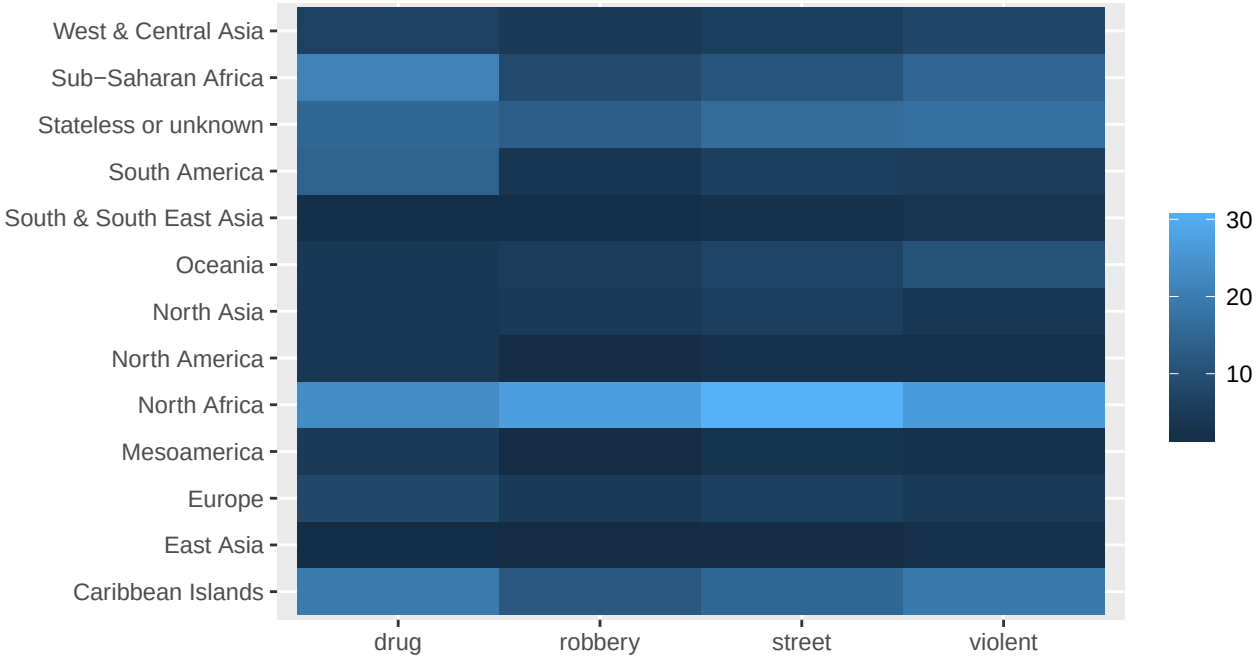
Nationality	Excluded Years	Population
Antigua and Barbuda	2012-2014	18
Holy See (Vatican City State)	2012-2015	2
Macao	2015-2016	10
Marshall Islands	2014-2015	1
Micronesia (Federated States of)	2013-2015	1
Monaco	2012-2016	13
Nauru	2012-2015	1
Palau	2013-2016	5
Saint Kitts and Nevis	2012-2016	10
San Marino	2012	15
Solomon Islands	2012-2016	5
Timor-Leste	2013-2016	2
Tuvalu	2015	18
Vanuatu	2012-2016	9

I first sought to get a sense of the prevalence of different types of crimes across time, independent of nationality group. The results can be seen in the graph on the next page. Violent crime (whether public or private), street crime (including street violence, robbery and theft, and property damage), and drug crimes are the most common categories in the dataset, at approximately 6 instances per 1000 people. The lion's share of the street crimes are robberies and thefts, which occur at a rate of approximately 4.7 instances per 1000 people. Rapes, murders, cybercrimes, and economic or white collar crimes occur less frequently, all at rates below 1 instance per 1000 people.

While the comparatively high rate of violent crime may be somewhat surprising, it is important to note that the PKS dataset only includes those crimes for which a suspect has been identified and arrested. This is much more likely to be the case for violent crime than for muggings or property damage, for example, both because the more personal nature of violent crime often makes it easy to identify a suspect, and due to greater police effort in identifying suspects for these more "serious" crimes.

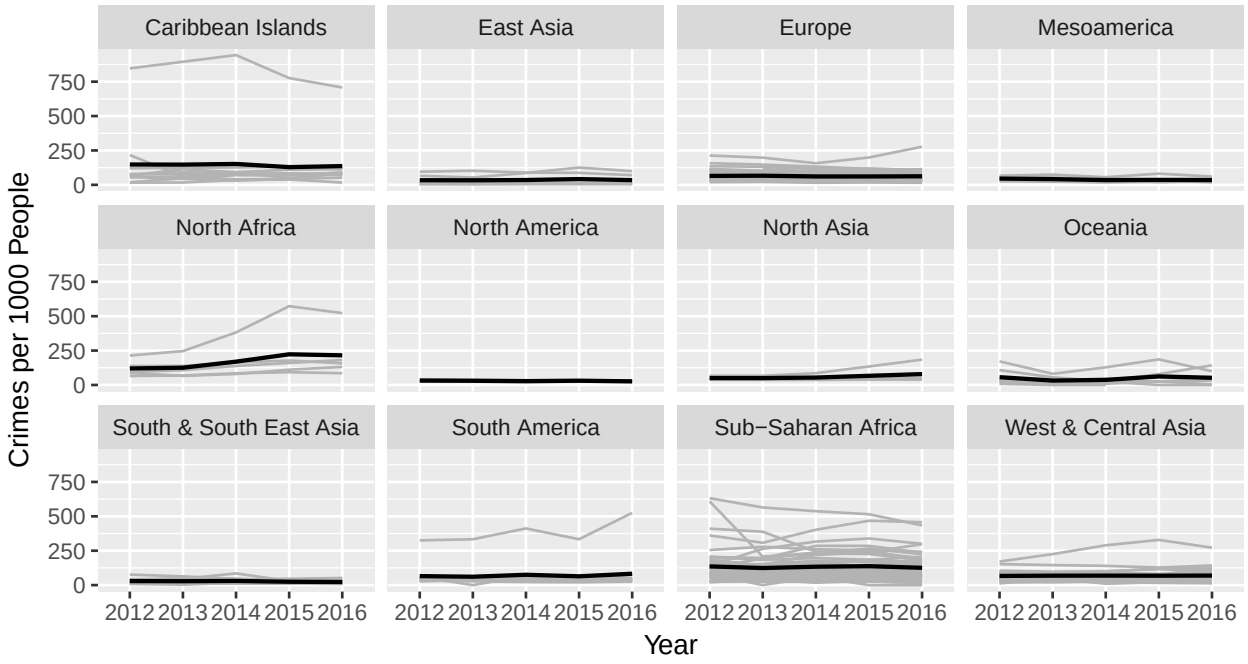


I next examined regional differences in the most frequently-occurring types of crimes. As can be seen from the heat map below, North Africans (including Algerians, Libyans, Moroccans, and Tunisians) have the highest rates of all four frequent types of crimes, while East Asians generally have the lowest rates. On the whole, there is little evidence of regional specialization in different types of crimes, with the exception of drug crimes among South Americans and sub-Saharan Africans.

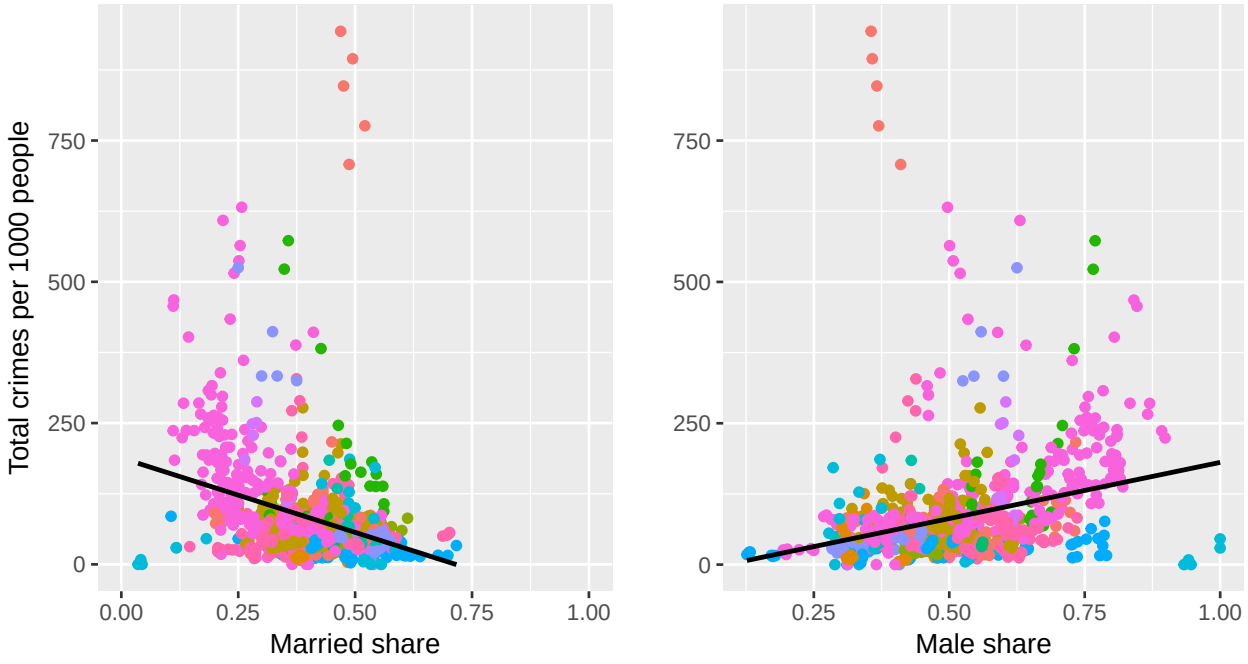


To further explore what might be driving such regional differences, I plotted the total number of crimes commited by members of each nationality group over time, faceted into separate graphs for each region (see below). This revealed that the crime rates for certain regional groups were likely affected by outliers: Dominica for the Caribbean, Algeria for North Africa, Suriname for South

America, and Georgia for West & Central Asia. Sub-Saharan Africa also stands out in this graph for its wide range of crime rates, likely due to the large number of states contained in this region.



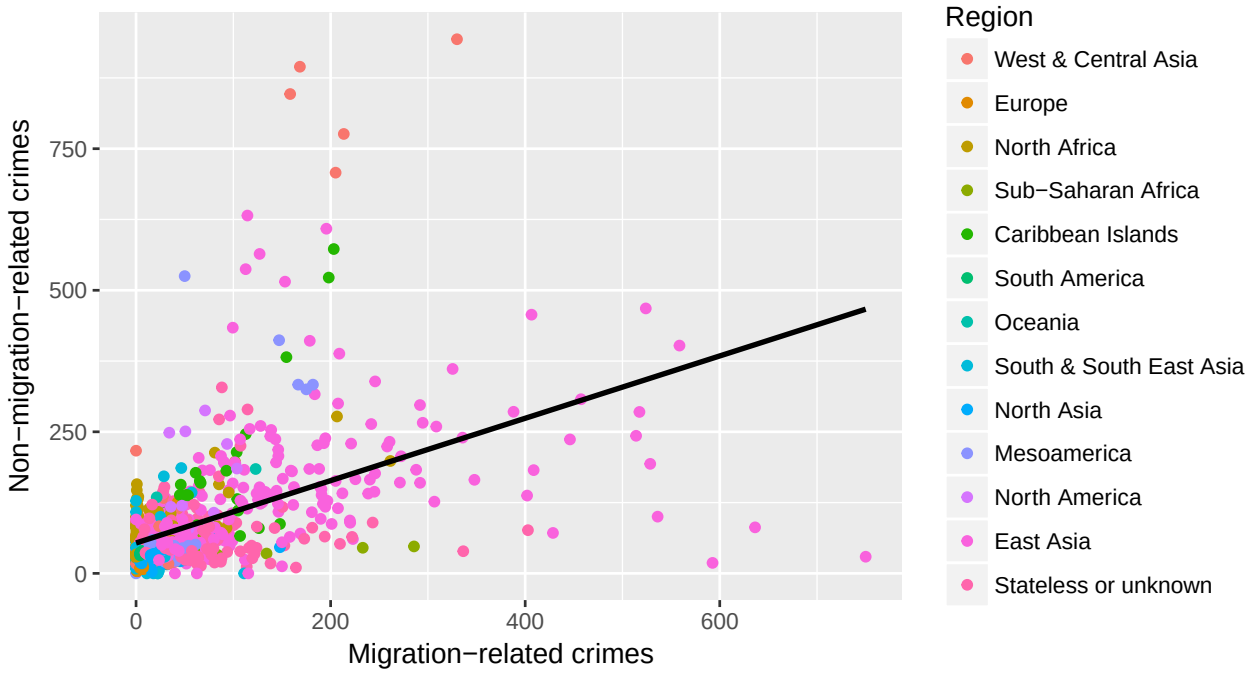
I next plotted the total number of crimes committed by each nationality group against the population-level control variables for gender and marital status (the different colors correspond to different regions; see legend of the graph on the next page). For easier interpretation, these were simplified to the share of residents who are male and share of residents who are married. Previous sociological analyses of crime indicate that these two factors are important predictors of criminal behavior.



This was no different in the current dataset, as can be seen from the plots above. Nationality groups with a lower share of married people commit substantially more crimes than nationality groups

with a higher share of married people, while the reverse is true of nationality groups with lower and higher shares of males. In addition, the nationality groups in the dataset differ vastly in terms of both gender and marital status, indicating the importance of controlling for these variables in statistical analyses.

As discussed above, I was not able to download population-level data on immigration status for each nationality group due to a paywall on the Destatis website, and thus could not test the relationship between asylum seeking and crime directly. However, I was able to explore this relationship with a proxy variable: the number of migration-related crimes (illegal residency and asylum violations) committed by members of each nationality group in the PKS dataset.



As can be seen in the graph above, there is a substantial positive correlation between migration-related crimes and non-migration-related crimes in the current dataset. However, this correlation should be interpreted with caution due to the likely non-independence of the two indicators. Police who arrest a criminal suspect for a non-migration-related crime are likely to charge him or her with a migration-related crime as well, if applicable. Thus, nationality groups who commit non-migration-related crimes at higher rates should be more likely to appear in the dataset for migration-related crimes than nationality groups with just as many migration-related offenses in reality but fewer non-migration-related arrests.

Statistical Analyses

I next ran several regressions to test my major hypotheses, which can be stated more formally as follows:

1. Nationality groups with higher rates of migration-related crimes will also have higher rates of non-migration-related crimes.
2. Crime rates after 2015 will be significantly higher than crime rates before 2015 due to the "uncontrolled" immigration of potentially criminal asylum seekers in 2015.

I included male share and married share in all models as control variables, and also tested their interaction to see whether it is specifically *unmarried males* who contribute to higher crime rates.

OLS Models

In Models 1-3, seen in Table 2 on the next page, I ignored the panel structure of the dataset and conducted pooled cross-sectional regressions. Dummy variables for each panel wave (except 2012, which served as baseline) were included in order to test whether a significant increase in crime rates occurred after 2015's heavy inflow of asylum seekers. This model specification was selected over a simple dummy variable for pre-/post-2015 in order to control for other potential annual effects, such as changes in the underlying data distribution over time.

Model 1, which includes controls for gender and marital status but not for region, explained roughly 25% of the variance in crime rates (R^2 = 0.255). The coefficient on migration crimes is positive and significant (b = 0.499, $p < .01$), with two additional migration-related crimes per 1000 residents associated with one additional non-migration-related crime. However, none of the year dummies are statistically significant, indicating that crime rates do not appear to have changed much over time. The coefficient on married is highly significant and in the expected direction (b = -92.099, $p < .01$): Moving from a 0% marriage rate to a 100% marriage rate is associated with a decline of 92 crimes per 1000 people. The coefficient on male, on the other hand, is not statistically significant, although it is in the expected direction.

In Model 2, I again ran a pooled cross-sectional model, but included an interaction term between married share and male share to test whether the effect of marriage differs among males and females. However, the interaction term was not statistically significant ($p > .10$); thus, I did not include interaction terms in subsequent models.

Next, I included regional dummy variables in the model in order to test whether migrants from particular world regions have higher crime rates than others. This model has not been shown due to space considerations. However, North Africans, Caribbeans, and people who are stateless or whose nationality is unknown all significantly differed from the baseline. They were thus included in a more limited regional model (Model 3 in the table above). These three regions are associated with substantial higher crimes rates, between 80-90 additional crimes per 1000 people. The fact that North Africans, but not people from West & Central Asia (including the rest of the Middle East), and Caribbeans, but not sub-Saharan Africans, have elevated crime rates indicates that these rates cannot be explained by discriminatory policing alone. It is not simply being Arab or black that leads to higher reported crime rates, but rather factors specific to Caribbean and North African migrants.

The strong positive coefficient on migration crimes was quite robust to model specification in the OLS regressions, barely shifting from Model 1 to Model 3. However, as discussed above, at least part of this relationship is likely due to bias in the way migration-related-crimes are tabulated.

Fixed effects model

In Model 4, seen on the right column of Table 3, I ran a model with time and nationality fixed effects in order to take into account the underlying panel data structure. Fixed effects models are also known as "within" estimators, as they calculate the effect of each independent variable on the

Table 2: OLS Regression Results

	Dependent variable:		
	Crimes per 1000 people		
	(1)	(2)	(3)
2013	−9.245	−9.236	−9.214
	(8.953)	(8.956)	(8.598)
2014	−8.848	−8.891	−9.034
	(8.976)	(8.980)	(8.620)
2015	−14.402	−14.470	−15.285*
	(9.010)	(9.015)	(8.654)
2016	−9.684	−9.615	−10.621
	(9.047)	(9.052)	(8.690)
Married share	−92.004***	−131.329	−137.631***
	(30.262)	(84.461)	(29.825)
Male share	12.848	−16.016	−23.509
	(25.183)	(63.122)	(24.846)
Migration crimes	0.499***	0.504***	0.502***
	(0.037)	(0.039)	(0.036)
Male x Married Share		73.947	
		(148.269)	
North Africa			89.562***
			(17.338)
Caribbean			80.153***
			(11.481)
Stateless or unknown			82.411***
			(26.360)
Constant	94.496***	110.084***	123.182***
	(21.630)	(38.014)	(21.313)
Observations	919	919	919
R^2	0.255	0.255	0.315
Adjusted R^2	0.249	0.249	0.308
Residual Std. Error	85.941 (df = 911)	85.976 (df = 910)	82.539 (df = 908)
F Statistic	44.548*** (df = 7; 911)	38.978*** (df = 8; 910)	41.769*** (df = 10; 908)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Fixed Effects Model

	Dependent variable:	
	Crimes per 1000 people	
	OLS	*panel linear*
2013	−9.245	−4.836
	(8.953)	(3.005)
2014	−8.848	−1.704
	(8.976)	(3.057)
2015	−14.402	−2.277
	(9.010)	(3.187)
2016	−9.684	−6.605**
	(9.047)	(3.278)
Married share	−92.004***	−206.098***
	(30.262)	(49.917)
Male share	12.848	−5.009
	(25.183)	(47.996)
Migration crimes	0.499***	−0.029
	(0.037)	(0.027)
Constant	94.496***	
	(21.630)	
Observations	919	919
R^2	0.255	0.030
Adjusted R^2	0.249	−0.229
Residual Std. Error	85.941 (df = 911)	
F Statistic	44.548*** (df = 7; 911)	3.162*** (df = 7; 725)
Note:		*p<0.1; **p<0.05; ***p<0.01

dependent variable within each panel unit (i.e. nationality group). This has the effect of controlling for all time-invariant cross-sectional variation, whether it has been explicitly included in the model or not. Thus, fixed effects models can eliminate some forms of omitted variable bias.

The coefficient on migration crimes is no longer significant in the fixed effects model and even has the opposite sign. This means that while the number of migration-related crimes is a strong predictor of non-migration-related crimes at the between-nationalities level, it is not significantly associated with changes in non-migration-related crimes within each nationality over time. This makes intuitive sense, as changes in the number of people of a given nationality arrested for illegal residency or asylum violations over time are likely due just as much to luck or policing priorities as they are to actual changes in the number of undocumented migrants.

The year dummy for 2016 has become significantly negative in the fixed-effects model. This serves as a sound rejection of Hypothesis 2, as crime fell in 2016 despite the arrival of nearly 1 million asylum seekers over the previous year. However, it is not sufficient evidence to conclude that the newly-arrived asylum seekers had a *negative* effect on crime, as a number of other factors could have led to the decline in 2016. This will need to be further explored after including data for more control variables.

Conclusion

In this paper, I conducted initial data cleaning and analysis of German Police Criminal Statistics (PKS) data on criminal suspects by nationality for the years 2012-2016, as well as a few control datasets. While a number of problems remain, mostly due to a lack of properly specified control variables, the results provide first indications of the relevance of my research topic and suggest avenues for further analysis. Specifically, there appears to be a strong positive link between illegal migration and crime - although this could well be reduced or eliminated with a better measure of illegal migration - as well as significant regional differences.

On the other hand, I found no support for the hypothesis that uncontrolled migration after 2015 led to an increase in crime. However, here again model specification issues are a cause for concern, as I would theoretically expect only an increase in crimes among nationality groups that experienced an increase in asylum seekers in 2015. This would require testing an interaction between year and number of asylum seekers. However, I elected not to test such a model in this paper because my proxy for asylum seeking - arrests for illegal entry and asylum violations - does not adequately measure the concept I had hoped to test.