Order in Chaos: Understanding Criminal Patterns in Germany

Rojan Abolhassani [*1] **Weronika Jaśkowiak** [*2] **Dóra Molnár** [*3] **Balázs Szabados** [*4]

Abstract

Crime prevention and criminal justice are shared concerns in today's society, driving the need for a comprehensive understanding of criminal patterns. In this project, we analyze datasets provided by the *Bundeskriminalamt* containing police crime statistics from the years 2002-2022. Our analysis includes two approaches for predictions for future case numbers as well as gaining insight in the most frequent offences, the age distribution of suspects, and the tendencies in clearance rates throughout time. We aim to enhance understanding of underlying reasons and correlations using various statistical methods and visualizations.

1. Introduction

In contemporary society, increasing concerns about crime prevention and upholding justice highlight the need for a deep understanding of criminal patterns. However, with numerous influencing factors, determining trends in criminal data is a considerable challenge. Our study focuses on police crime statistics from 2002 to 2022, aiming to extract key insights and forecast future number of offences.

During the analysis, we focused on two main approaches. In the first part, we elaborated on predicting the total number of offences for the upcoming years. We fitted various regression models (Section 3.1) to the data spanning from 2002 to 2022. We compared the methods based on their error rates on randomly chosen test years. Additionally, we trained an autoregressive model to explore the time series perspective of the data (Section 3.1). The predictions from the two methods are detailed in Section 4.1. In the second part we investigated two important aspects of criminal activity in Germany: age distribution among suspects (Section 3.2) and effectiveness of the police (clearance rates) (Section 4.2). The detailed results of our analysis and visualizations are included in Section 4. In Section 5 we present a sociological background that might have significantly influenced changes in criminal trends. All details of our work is also available on GitHub.

[*]Equal contribution [1]Matrikelnummer 6665489, rojan.abolhassani@student.uni-tuebingen.de, MSc Machine Learning [2]Matrikelnummer 6761084, weronika.jaskowiak@student.uni-tuebingen.de, MSc Bioinformatics [3]Matrikelnummer 6728487, dora.molnar@student.uni-tuebingen.de, MSc Machine Learning [4]Matrikelnummer 6727211, balazs.szabados@student.uni-tuebingen.de, MSc Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

2. Data

For this project, we used multiple datasets consisting of PDF and Excel files from the website of the *Bundeskriminalamt*, covering police crime statistics spanning from 2002 to 2022. Datasets from 2002 to 2013 are stored in PDF format and present a concise overview of crime trends. Among other features, they include the offence key, offence category name, number of recorded cases, change in the number of recorded cases compared to the previous year and clearance rates in percentage. We extracted this data into Excel files (Cases20**.xlsx).

Data between 2014 and 2022 are stored in Excel files and provide more detailed information. The data are organized into four tables:
Cases Basic Table - most importantly includes the offence key, offence category name, number of recorded cases, distribution of crime scenes grouped by the number of inhabitants in the area, clearance rate in percentage and the number of non-German suspects.
Cases Development - includes the offence key, offence category name, number of recorded cases, its change in percentage, and the clearance rate in percentage from the actual and previous year.
Suspects - provides information on suspects per offence, including their sex and age category.
Victims - provides information on victims per offence, including their sex and age category.

The listed offences are assigned to keys. It is possible to distinguish main key groups based on characteristics shared by similar offences. For instance, the key '500000' correspond to offences related to property and forgery, but distinctions exist within this group: '530000' stands for misappropriation and '540000' for document forgery. Within the main key groups, there are also smaller categories that aggregate specific types of offences.

Depending on the conducted analysis, we utilized both main key groups and more detailed ones related to specific types of crimes.

To grasp a first understanding of the data, Figure 1 below provides an overview of the total number of cases, suspects, and victims from the examined years.

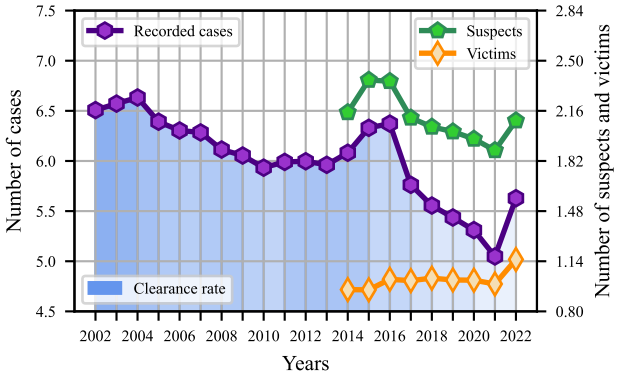

Figure 1. The total number of recorded cases, suspects and victims each year in millions. The left y-axis represents recorded cases, while the right y-axis corresponds to both suspects and victims. The shading of each year's column indicates its average clearance rate, with darker colors representing lower rates and brighter colors reflecting higher rates. Note that all clearance rates fall within the range of 50-60%.

3. Methods

3.1. Predicting total case numbers for the following years

Our work involves two different approaches for predicting the total offence numbers in 2023 and beyond. Firstly, we fitted various regression models to the data between 2002 and 2022 using the Scikit-learn library of Python, including linear regression, kernel ridge regression, random forest regression, and decision tree regression. To assess the accuracy of these models, we randomly selected four test years and computed the R^2 and mean squared error on these data points. We also fitted polynomials with higher degrees to our data. Although the test errors of this method generally outperformed those of the previously mentioned models, it is not the most suitable technique for predicting the number of upcoming years, as further detailed later.

Building upon our regression analyses, we further explored the time series aspect of the dataset to forecast the total offence numbers for the next five years. We used an autoregressive model for our time series analysis. To address non-stationarity in the data, we applied first-order differencing. Influenced by an examination of the partial autocorrelation function, we chose an autoregressive component of order 5. The decision not to include a moving average component was based on insights from the autocorrelation function. We then trained and tested the model using the mean absolute percentage error, given the large scale of the total offences. Further results of these methods are discussed in Section 4.1.

3.2. Age distribution among suspects

For the analysis of suspect age groups, we created a permutation test to determine whether the distribution of the number of suspects over the age groups is the same for males and females.

We also attempted to examine if suspect ages are normally distributed through calculating its parameters and fitting different probability density functions to the data.

4. Results

4.1. Predicting total case numbers for the following years

Due to the relatively small dataset, the performance of the fitted models varies significantly, mostly depending on the randomly chosen test years. Fitting polynomials with higher degrees usually resulted in lower error rates; however, due to the nature of these functions, they predict unrealistically high (or low) values for years outside the fitted region. Because of this, we continued our prediction with four models.

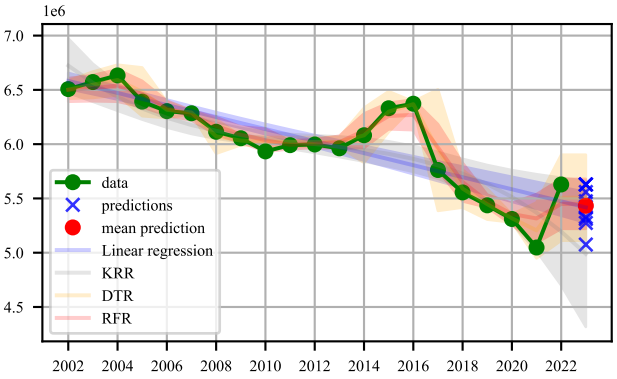

Figure 2. Performance of linear regression, kernel ridge regression (KRR), decision tree regression (DTR) and random forest regression (RFR). For each method the standard deviation and the mean are visualized based on 10 rounds of training with 4 randomly chosen test years. The final prediction for 2023 is the mean of the 10 values given by the model with the least mean squared error in each round.

Throughout the experiments, the best accuracy was achieved by the random forest and decision tree regressors. To obtain the final prediction for 2023, we fitted our models in 10 rounds and calculated the mean of the predictions given by

the model with the lowest mean squared error in each round. The final prediction using this method was: 5,436,990.

Using the autoregressive model from Section 3.1, the calculated error rate stands at 7.36%, which, given the inherent variability in crime data, is relatively low. We then forecasted five steps ahead, spanning from 2023 to 2027, extending our analysis into the future. Specifically, our model predicts a total of 5,918,203 offences for the year 2023. Additionally, we complemented this analysis with generated plots that visually illustrate the steps of the method and the forecasted total numbers of offences for the specified period.

4.2. Age distribution among suspects

We formulated a hypothesis suggesting that the distribution of the number of suspects across age groups remains consistent between males and females. To test this hypothesis, we created a permutation test utilizing suspect data from the year 2022. The two samples for the test are the ages of male and female suspects. As the data only provides age groups and not exact ages, we generate the ages for the samples using the mean of each age group. In addition, we only use a thousandth of the suspects to be able to calculate a sufficient amount of permutations. As our test statistic, we chose the absolute difference of the sample means. Generating a million permutations, we obtained a p-value ≈ 0.155. Since by calculating more permutations we can only increase this value, we retain our hypothesis.

Another assumption we had was that the age of suspects might be normally distributed. We suggested that choosing the mean around 37 years, it would be reasonable to assume that it is equally likely that a suspect's age is 20 or 54 years. The following plot disproves this hypothesis: although the mean of the distribution is close to 37, it initially grows quicker than the Gaussian with the parameters derived from the data and, as a result, decays more slowly.

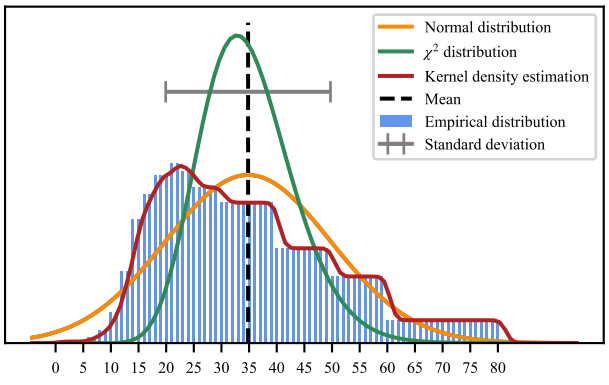

Figure 3. The distribution of suspect ages. After calculating the mean and standard deviation of the empirical distribution, we plotted a Gaussian and a χ^2 probability density function for comparison. Additionally, we computed a kernel density estimation.

4.3. Clearance rates

Clearance rate is one of the key metrics for assessing police efficiency. Based on the data, two important tendencies can be observed when looking at these percentages. Firstly, as shown in Figure 1 and Figure 4, the average rate of crimes solved by the police has an increasing value between the years 2002 and 2022. However, we have to note that the total number of offences shows a decreasing trend over the same period. Secondly, a strong correlation can be detected between the clearance rates and the types of crimes. Figure 4 visualizes this observation. We can see how offences in the category 'theft and burglary' have the lowest rates throughout the examined years while, for instance, the majority of 'offences against personal freedom' are solved by the police.

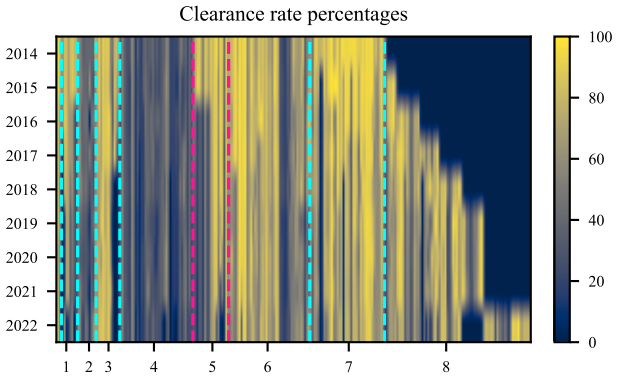

Figure 4. The clearance rate percentage of offences each year. Each column represents a single offence. Dashed lines serve to group similar crimes together, white-collar crimes are highlighted with purple. 1 : Offences against life and sexual self-determination, 2 : Robbery, 3 : Offences against personal freedom, 4 : Theft and burglary, 5 : White-collar crimes, 6 : Other criminal offences, 7 : Supplementary criminal laws, 8 : Aggregate keys. In the last category many of the offences were not recorded in earlier years.

4.4. Tendencies in the number of recorded crimes from 2002 to 2022

The analysis of the total recorded cases revealed significant changes in separate periods, notably the highest number of recorded cases in 2004, followed by an increase in the years 2014-2015 compared to the previous years and a substantial decrease after 2016. A deeper insight into the data from 2004 and 2005 revealed that hostile sentiment among the public was on the rise, leading to the highest number of crimes in categories such as thefts, street crime, and fraud. The noticeable increase in 2014 and 2015 was especially prominent among non-German suspects (see Figure 5), but we have to note that during this period, offences against foreigners are also at their highest across all years. After 2016, there is a monotonic decline in the number of recorded

cases, which is only interrupted by the last recorded value. Some of the possible underlying reasons and influences are discussed in Section 5.

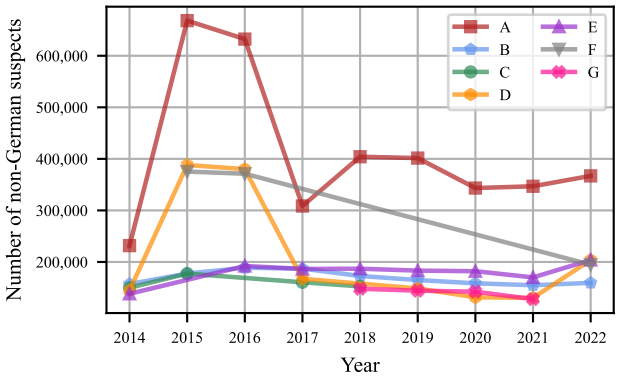

Figure 5. Top 5 offences among non-German suspects in years 2014-2022. A: Unauthorised stay, B: Property and forgery offences, C: Total thefts, D: Offences against the Residence Act, the Asylum Act and the Freedom of Movement Act E.U., E: Offences involving brutality and crimes against personal freedom, F: Illegal entry, G: Bodily injury.

5. Discussion & Conclusion

In this project, we analyzed various datasets describing criminal activity in Germany between the years 2002 and 2022. Our work focused on predicting the total number of recorded cases for the upcoming years and examining different characteristics related to the recorded offences, such as the distribution of the age of suspects and the effectiveness of the police based on clearance rates. The main limitation of our work is the uncertainty surrounding the exact number of committed offences. Numerous factors may impact the willingness to report a crime, such as its severity and immediacy, the general trust in the work of the police, and the consequences of filing a report. Nevertheless, there are important societal-level influencing elements, including worldwide restrictions due to a pandemic (Liu et al., 2022).

Lastly, it's important to highlight some key social, legal, and political events that could have had a significant impact on the changing number of recorded crimes from 2002 to 2022. Starting in 2004, an economic crisis occurred. The government, led by Chancellor Gerhard Schröder, implemented labor market reforms known as the 'Hartz reforms' between 2003 and 2005 (Blömer et al., 2015; James, 2004). In 2014 and 2015, Germany faced an unprecedented surge of asylum seekers and migrants, with a staggering influx of over 890,000 individuals into the country. At that time, Chancellor Angela Merkel declared that Germany would welcome refugees without imposing any legal limits (Sola, 2015; Wassmer, 2019). Between 2013 and 2017, German authorities underwent substantial legislative changes, particularly in the realm of criminal law. The goal was to address a spectrum of issues, including corruption in both public and private sectors, hate crimes, stalking, telecommunications interception, criminal asset confiscation, and corporate sanctions law (Wassmer, 2019). In 2020, the outbreak of the global COVID-19 pandemic could have led to a decrease in the number of recorded cases (Liu et al., 2022).

Contribution Statement

Rojan Abolhassani made significant contributions to the prediction of case numbers and the analysis of clearance rates. Weronika Jaśkowiak focused on thoroughly examining various social aspects related to the dataset. Dóra Molnár played a role in data preparation, case number prediction, and the comprehensive analysis of clearance rates and age groups. Balázs Szabados actively participated in data preparation, conducted in-depth age group analysis, and contributed to the visualization of key insights. All team members conducted foundational data analysis and jointly co-authored this report.

References

Blömer, M. J., Dolls, M., Fuest, C., Löoffler, M., and Peichl, A. German public finances through the financial crisis. *Fiscal Studies*, 2015.

James, K. Germany 2004: The road to reform. *Deutsche Welle*, 2004.

Liu, L., Chang, J., Long, D., and Liu, H. Analyzing the impact of covid-19 lockdowns on violent crime. *Int J Environ Res Public Health*, 2022.

Sola, A. The impact of the 2015 refugee crisis on concerns about immigration in germany. *SOEPpapers on Multidisciplinary Panel Data Research at DIW Berlin*, 2015.

Wassmer, M. The latest criminal law reforms in the general and special part of the german criminal code. *Pravo. Zhurnal Vysshey shkoly ekonomiki*, 2019.