Temporal and Spatial Exploration of Digital Crime in Germany

Tim Gerne [*1] **Tim Haigis** [*2] **Monica Janders** [*3] **Lukas Weber** [*4]

Abstract

A significant portion of our routine actions is frequently carried out in the digital realm. Consequently, we explore whether this shift is not only apparent in lawful contexts but also observable in criminal activities. To this end we source datasets from the German Federal Office of Criminal Investigation due their supposed reliability and public availability.

After assesing data integrity and sufficiency, our two-fold analysis reveals a linear increase in cybercrime cases and high concentration in city states. We also observe a strong negative correlation between digital and analog fraud for the last ten years and discuss limitations of governmental data.

1. Introduction

After examining the available data thoroughly, we refrain from our initial idea of relating major events in Germany to crime rate development as the German Federal Office of Criminal Investigation, known as the Bundeskriminalamt (BKA), does not provide datasets with sufficiently high spatio-temporal resolution. Instead we focus on long-term development of crimes and their regional differences.

We start by describing and examining the utilized datasets (Section 2). Then we segment our analysis into temporal and spatial aspects respectively, reflecting the structure of the data. We compare trends in cybercrime and conventional crime and attempt to analyze the development of computer fraud (Section 3). Cybercrime is defined by the BKA as: "Cybercrime comprises criminal offenses that are directed against the Internet, data networks, information technology systems and their data or which are committed using this information technology" (BKA)[1]. According to §263 of the German StGB "Fraud penalizes individuals who, with the intention of gaining an unlawful financial advantage for themselves or others, damage another person's wealth by inducing or maintaining a misconception through the presentation of false facts or the distortion or suppression of true facts." (stg, 2023)[1] Additionally, we highlight regions characterized by high digital crime rates and provide comparisons of relevant geographic normalization methods (Section 4). We conclude by discussing our findings and address limitations of the data (Section 5).

2. Data Assessment

Utilizing data from the BKA, we have access to various datasets on criminal cases in Germany since 1987, consisting of annually compiled files. Although digital files from 1953 are available, their scanned format requires unreasonable effort for conversion into tabular files. The data availability is increased since a 2012 system update by the BKA, including monthly case numbers for Germany and time series with annual overviews dating back to 1987. Another update adds an annual distribution of cases across federal states, districts and cities, enabling a spatial analysis. Depending on the offense and analysis type, usable data dates back to 1987 or 2012, with some analyses starting in 2016 due to legal changes.

For further readability we introduce columns of common tables containing relevant information for this project. It is important to note that the exact naming of columns differs in some tables as the labeling and overall format changes over the years.

The annual federal tables (BKA, 2022d) (BKA, 2022a) include the column "Erfasste Fälle" containing registered cases on a number of crimes. Individual categories of crimes are encoded with numerical keys along with numerical group keys which aggregate multiple categories into a single key. The columns related to "Tatortverteilung" are segmented into various ranges based on the population size of the areas where the crimes took place, serving as crucial data for our spatial analysis.

The monthly federal table (BKA, 2022b) offers columns with monthly data for each year using similar keys as the

[*]Equal contribution [1]Matrikelnummer 5465398, tim.gerne@student.uni-tuebingen.de, MSc Machine Learning [2]5481789, tim.haigis@student.uni-tuebingen.de, MSc Cognitive Science [3]Matrikelnummer 5570379, monica.janders@student.uni-tuebingen.de, MSc Machine Learning [4]Matrikelnummer 5406074, lukas2.weber@student.uni-tuebingen.de, MSc Computer Science.

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.

[1]Translated by the authors

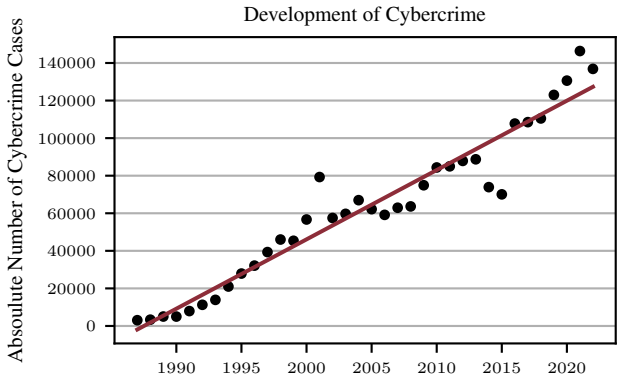

Figure 1. Development of cybercrime in Germany in absolute numbers. The method of least squares is used to fit a first degree polynomial with determination coefficient $R^2 = 0.930$.

federal base table. The "Tatzeit unbekannt" column includes cases without a specific month attribution, while the column "Anzahl Fälle insgesamt" contains the total number of cases in that year.
The annual state table (BKA, 2022c), compiled by the BKA, uses base tables from federal state police departments (LKA) like the annual federal table, grouping cases by federal state.

The PKS provides two tables containing regional information as well as sufficiently detailed crime keys: The national base table and the federal base table. The former includes columns where the number of cases are grouped by population size of the region in which they were recorded. These regions are distinctly categorized by population, with labels such as $<20k$, $20k$-$100k$, $100k$-$500k$ and $>500k$.
The PKS provides no clear definition on what is considered a region. Looking at city states where all cases are recorded in the $>500k$-category, we can place a lower bound on the granularity of regions, but without official documentation it is impossible to reliably map geographical regions to number of inhabitants. For examples and potential solutions see our GitHub repository (Gerne et al., 2024a).
The federal base tables avoid these issues altogether as they group cases by federal states. The downside of these tables is a lower data consistency, due to the the data being collected by each federal state's LKA independently.

3. Temporal Analysis

We first assess whether our different datasets with monthly and yearly crime cases match with regard to the absolute number of cases. We then conduct an analysis on the development of cybercrime and its subcategory, computer fraud, over the span of multiple years.

3.1. Data Consistency

To check for consistency of the data within the monthly federal table, we sum up the monthly cases and the cases of the column "Tatzeit unbekannt" for overall crimes and compared this sum with the column " Anzahl Fälle insgesamt". The values turn out to be identical for every year except for 2013 as there is no column "Tatzeit unbekannt". We assume that the missing column accounts for the higher value of "Anzahl Fälle insgesamt" in this year as the difference has the same order of magnitude as the values of "Tatzeit unbekannt" for the other years.
To ensure data consistency between the annual federal tables and the monthly federal tables we compare "Anzahl Fälle insgesamt" from the former with "Erfasste Fälle" from the latter. It is shown that the amount of cases in the annual federal table are consistently higher than the values from the monthly federal table. To find an explanation for these differences we subtract the number of attempted crimes from the overall crimes from the annual federal table. While this leads to a notable decrease in the number of total crimes it does not lead to the data of annually and monthly federal crimes being identical.
The annual federal table exhibits higher case numbers than the monthly federal table in certain years, while in other years, the relationship is reversed. For a visualization of the data discrepancy, please refer to Gerne et al. (2024b). The additionally provided information on the tables by the BKA do not lead to more insights on this discrepancy. For the scope of this project, it imposes a restriction on our ability to compare data from annually and monthly federal tables due to the unknown underlying differences in data collection.

3.2. Data Analysis

An initial analysis of the development of cybercrime cases shows an increase between 1987 and 2022, which can be approximated by a polynomial of first degree with a coefficient of determination of $R^2 = 0.930$, shown in figure 1. In our analysis, polynomial fittings of degrees two and three had just a slightly higher value for R^2 and show similarities to the linear trend. This development intrigued us to look for further interesting development of cybercrimes and how they compare to analog crimes. Due to the high difference in magnitude of conventional crime and cybercrime numbers, we decide to focus on a subset of crimes and corresponding cybercrimes: fraud and computer fraud.While cybercrime is only up to ~3% of overall crimes, computer fraud is up to ~15% of overall fraud. Therefore, a comparison of fraud and computer fraud seems more reasonable. For further insights we look at the case numbers of fraud and computer fraud on a monthly basis between 2012 and 2022.
While visualizing the monthly cases in figure 2, we encounter a change of keys indexing the cases of cybercrime. For the data between 2012 and 2015 the key 517500 is used

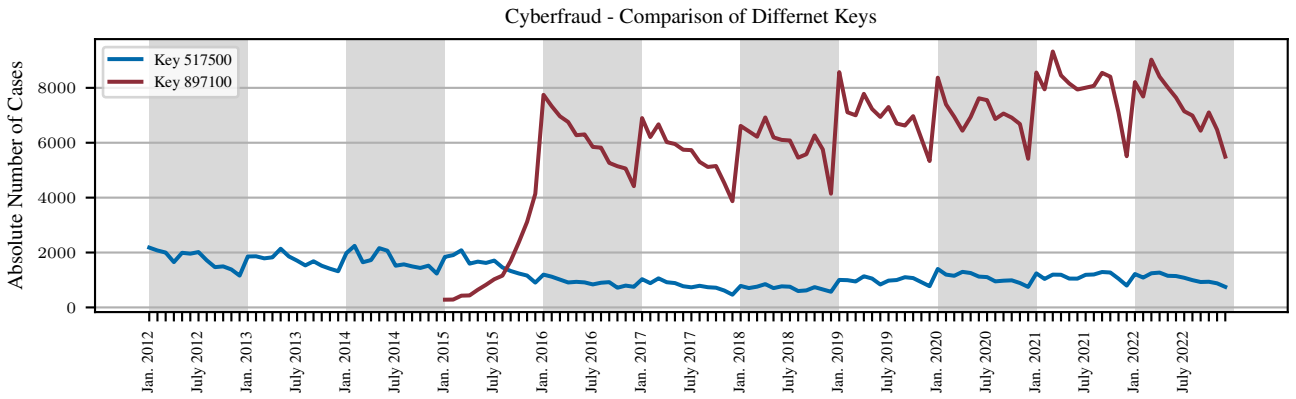

Figure 2. Comparing the different keys encoding computer fraud, it can be seen that the later used key 897100 presents higher numbers than the initial key 517500. The only exceptions are the first few month after the new key is first introduced. The numbers differ by several orders of magnitude.

for computer fraud and from 2016 onwards the same key is used for the crime computer fraud (other) ("Computerbetrug (sonstiger)"). Furthermore, from 2015 onwards the key 897100 is used to track computer fraud. This key is not used before in the PKS. There is an overlap of the two keys for computer fraud for the year 2015, leading to the question which key to use to track computer fraud. As the BKA does not provide any guidance, we plot the development of both keys. Initially there are lower values for the new key compared to the old key followed by a strong increase of the new key within the first year. The development of the initial key does not show any changes of magnitude over the years.

4. Spatial Analysis

After considering nation-wide data in the temporal analysis, we aim to dispel the implicit assumption of uniformly distributed crime rates with an exploratory spatial analysis. To this end we differentiate between regional differences at the expense of temporal scope and resolution, limiting ourselves to select yearly data. In the following we will first take a look at the available data, augment it with geographic information and compute two statistics for contrasted visualization.

The primary source for boundary data is based on the OpenStreetMap community (Opendatasoft et al., 2019) and extended with current information on population and area per federal state by Destatis (2023).

4.1. Spatial Methods

For the spatial data exploration we merge the LKA tables with the geographic data on federal states. In addition to the absolute number of cases provided by the table, we compute several relative statistics element-wise for each federal state:

The fraction of cybercrimes of all cases is calculated via

$$\%\ \text{cybercrime} := \frac{\#\ \text{all cases}}{\#\ \text{cybercrime cases}}.$$

For years where regional numbers of cases are not officially provided in the table, we normalize the absolute number of cybercimes by the population size with

$$\text{cybercrime per capita} := \frac{\#\ \text{cybercrime cases}}{\#\ \text{inhabitants}} \cdot 100,000.$$

The years 2016 and 2022 are selected for comparison to avoid issues with key changes and temporal anomalies based on findings from section 3.

4.2. Spatial Results

The spatial distribution of cybercrimes is split into multiple categories and compared over two years in figure 3. The first column shows data from 2016, while the second contains the most recent data from 2022. The general trend of overall increasing cybercrime rates is apparent in all shown statistics and correlates with the findings of the temporal analysis.
The first row shows the absolute number of cases, serving as baseline for further comparisons. Noteworthy is the overall high occurrence in North Rhine-Westphalia, Bavaria and Berlin. The second row highlights the relatively high cybercrime rate w.r.t. all crimes in Berlin and a strong increase in Saxony-Anhalt, Hamburg and Bremen. The third row places particular emphasis on the difference between city states and larger federal states, as the number of cybercrimes per capita supersedes the latter by multiple orders of magnitude. In line with the trends in cybercrime rate, Saxony-Anhalt, Hamburg and Bremen display significant increases during this six year period.

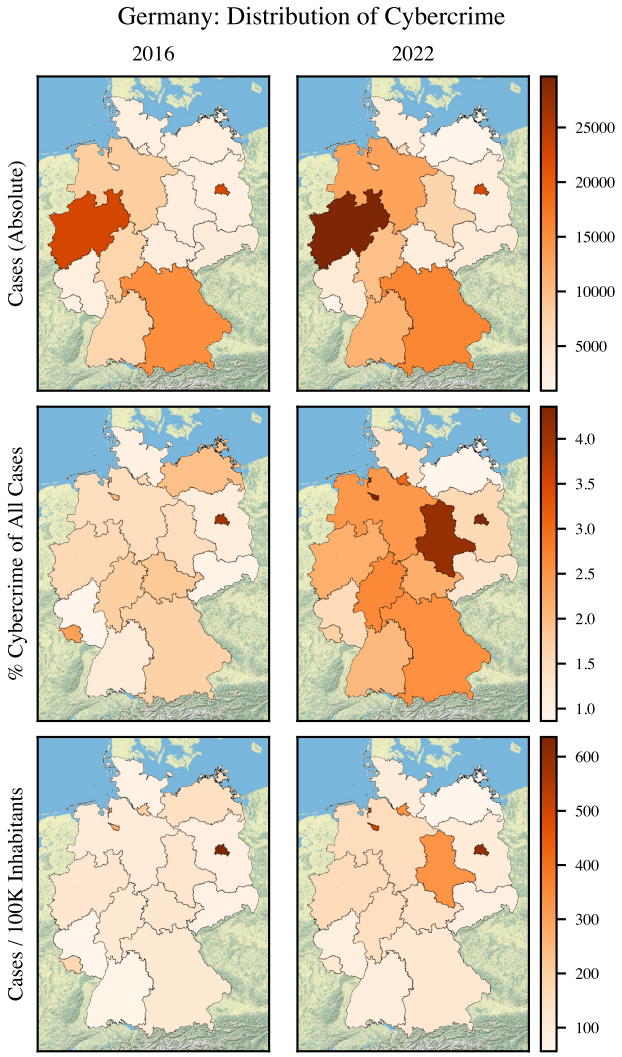

Figure 3. (*Top*) High number of cybercrime cases in North Rhine-Westphalia and Berlin along with general increase from 2016 (*Left*) to 2022 (*Right*). (*Middle*) Strong increase of cybercrime ratio of all crimes particularly in Saxony-Anhalt. (*Bottom*) Concentration of cybercrime in city states and Saxony-Anhalt defined by occurrence per 100k inhabitants.

5. Discussion

As there is little guidance provided by the BKA to help with the occurrence of key changes for computer fraud, we contacted an expert in this field, Prof. Dr. Jörg Kinzig, who serves as the Director of the Institute of Criminology at the Chair of Criminology, Criminal and Sanctions Law at the University of Tübingen. He redirected us to Benedikt Iberl, a doctoral candidate at the same institute. Among other insights, he informed us that as of January 2016, cybercrime is recorded using new underlying definitions, for what qualifies as cybercrime. Furthermore, based on the data he inferred the possibility that not only has the official definition of computer fraud changed, but also the internal categorization used by the BKA. He sent us an extract from the PKS stating that comparisons of any kind of fraud before and after 2016 are only possible to limited extent.

We fit the development of the absolute numbers of cybercrimes with a polynomial of first degree and compute $R^2 = 0.930$. This means it achieves a high variance explanation without overfitting. From further exploration in Gerne et al. (2024b), we see that the development of cybercrime in absolute numbers (as shown in figure 2) does not differ much from relative crime numbers, with nearly the same R^2 value for their linear approximation.

We see that the data points of 2014 and 2015 in figure 1 are far below the approximated line. The decrease in 2014 is explained in Köppen (2015) which states that only cases are recorded, where the offense most likely took place in Germany. The reason for the increase in 2016 is based on the increase of computer fraud in this year. The graph of all crimes does not show such an obvious polynomial correlation, but there is a visible downwards trend that there is a downwards trend since about 20 years.

To find an approximation for the development of computer fraud cases is more difficult due to the legal changes and the corresponding discontinuity in the graph in 2016. We observed a negative correlation between the fraud cases and the computer fraud cases since 2016.

While North Rhine-Westphalia stands out in terms of absolute crime rates, it is also the largest federal state by population and thus insignificant when considering relative rates. The reason for the observed concentration of cybercrime in city states and Berlin in particular remains unclear. We suspect this to be an artifact of the way reports are centrally collected, but a brief literature review revealed no additional insights. However the high cybercrime rate in Saxony-Anhalt does not seem to be an artifact and might be worth investigating more closely in future annual reports.

Although we analyzed the trends mentioned above, we want to point out that the reliability of the provided data is not as high as we hoped in the beginning. Köppen (2015) already referred to unreliable documentation and quoted a press release of the Federation of German Detectives: “Citizens get the impression [when the PKS is published] that it is a statistical reflection of reality. Unfortunately, we are a long way from this. In fact, the statistics provide at best a very incomplete record of the work of the criminal investigation department.” [1]. We hope these limitations inspire change in the overall approach to publishing and maintaining data of societal interest.

Contribution Statement

Tim Gerne and Lukas Weber did temporal analysis and data assessment. Monica Janders did the mathematical analysis and plots. Tim Haigis anlyzed and imported spatial data.

References

Strafgesetzbuch. Deutscher Taschenbuch Verlag, 2023. §263, German Criminal Code. Accessed: 2024-01-25.

BKA. Polizeiliche Kriminalstatistik (PKS) Jahrbuch 2014. PKS. Accessed: 2024-01-22.

BKA. T01-Bund-Fälle table 2022. PKS, 2022a. Accessed: 2023-12-06.

BKA. T08-Bund-Tatzeit table 2022. PKS, 2022b. Accessed: 2023-12-06.

BKA. T01-Land-Fälle table 2022. PKS, 2022c. Accessed: 2023-12-06.

BKA. Polizeiliche Kriminalstatistik. PKS, 2022d. Accessed: 2023-12-06.

Destatis. Bundesländer mit Hauptstädten nach Fläche, Bevölkerung und Bevölkerungsdichte am 31.12.2022. Publikation, 2023. Accessed: 2024-01-25.

Gerne, T., Haigis, T., Janders, M., and Weber, L. Spatial Analysis. GitHub repository, 2024a.

Gerne, T., Haigis, T., Janders, M., and Weber, L. Temporal Analysis. GitHub repository, 2024b.

Köppen, H. Entwicklung der Cybercrime im Jahr 2014. *Datenschutz und Datensicherheit*, 39:759–761, 2015.

Opendatasoft, OpenStreetMap, suche-postleitzahl.org, and BKG. Postleitzahlen - Germany. Dataset, 2019. Accessed: 2023-12-22.