A Data Science Approach to Spatio-Temporal Crime Modeling: Data, Models, and Applications

DISS. ETH NO. 25961

A Data Science Approach to Spatio-Temporal Crime Modeling: Data, Models, and Applications

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH Zurich

(Dr. sc. ETH Zurich)

presented by

Cristina Kadar

Master of Science

Elite-Masterstudiengang Software Engineering

Universität Augsburg, Technische Universität München,

Ludwigs-Maximilians-Universität München

born on *28.06.1984*

citizen of *Romania*

accepted on the recommendation of

Prof. Dr. Elgar Fleisch
Prof. Dr. Stefan Feuerriegel

2019

Abstract

Crime is pervasive in everyday life, leading to considerable negative effects for individuals and society as a whole, including financial losses, physical harm, psychological distress, and a general reduced quality of life. Furthermore, crime occurs in all types of environments: from highly populated to less populated areas, and from developing countries to the most developed countries in the world. As such, it is of the utmost importance for research to advance our understanding of crime and to support policy makers, as well as current decision-makers in developing and implementing effective measures for preventing crime.

Crime is not random, but the result of a complex interaction between offenders, targets, and the environment. Several types of crimes, including both property crimes and violent crimes, have been shown to reveal spatio-temporal patterns. Long-term crime patterns can persist for years and even decades, while increased short-term risk has been found over a horizon period of days to weeks. Hence, both long- and short-term crime prediction models have theoretical and operational value.

The advent of urban and social big data, paired by constant innovation in machine learning research, has fueled the development of novel research in the emerging fields of computational social science and urban computing. These studies have lead to better understanding of human behavior and developed analytical solutions to urban challenges. This interdisciplinary thesis taps into this potential and explores research opportunities emerging from the application or the development of state-of-the-art data science techniques on ubiquitous data. It presents several studies aiming to advance the study of spatio-temporal crime concentrations along three dimensions: data, models, and applications. The benefits for the study of crime are two-fold, as this thesis: (1) proposes innovations in terms of spatio-temporal data and methods for long- and short-term crime prediction and their applications and implications for both researchers and practitioners, and (2) puts forward a novel application of data from location-based services towards a very granular and scalable operationalization of crime theory that advances our general understanding of the crime phenomenon.

Several crime theories, including the social disorganization theory, have proposed that crime concentrations can be explained to a certain extent by the social attributes of the neighborhood (such as income levels, residential

instability, and ethnic heterogeneity). These theories have been extensively validated in empirical studies utilizing census statistics describing the resident population of the neighborhood. Other theories, such as the crime pattern theory, have proposed a link between crime and the attributes and the activities of the ambient population in the neighborhood. Yet, until now, the corresponding empirical studies have had limited data to work with. Hence, the first study presented in this thesis evaluates different urban data sources capturing the activities of the ambient population in a neighborhood for integration in long-term crime prediction models. Aiming at predicting yearly neighborhood crime counts, the case study leverages several tree-based ensemble models from machine learning on the large and heterogeneous resulting set of features based on Foursquare venues and checkins, subway exits/entries, and taxi rides. The model specifications including human activity features always improve upon the baselines leveraging only census features for all crime types and across different evaluations. For instance, the R^2 metric increased by 30 percentage points when predicting crime in unseen neighborhoods (geographical test set) and by 7 percentage points when predicting crime in the following year (temporal test set). The predictive gain from adding the novel features varies across crime types: such features bring the biggest boost in case of grand larcenies, whereas assaults are already well predicted by the census features. Notably, the full models incorporating features inferred from Foursquare data deliver the highest prognostic capacity compared to the models incorporating features derived from taxi or subway data. The highest improvement brought by the human activity features has been observed in busy neighborhoods of the city, such recreational and shopping areas. Besides advancing the urban computing literature, these results offer valuable insights for those responsible for urban development.

Prior studies, including the first study of this thesis, are limited to aggregated statistics of visitors when assessing the ambient population in an area. Because of that, they neglect the temporal dynamics of individual human movements. As a remedy, the second study of this thesis presents the first work which studies the ability of granular human mobility in describing and predicting spatio-temporal crime concentrations. For this purpose, the study proposes the use of data from location-based services, specifically Foursquare. This type of data consists of individual transitions and offers the

possibility to distinguish between different types mobility flows: (1) incoming or outgoing from a neighborhood, (2) self-looping within it, or (3) transitions where people only pass through the neighborhood. As such, the first two types of mobility flows model the concept of routine activity nodes from crime pattern theory, i. e., locations where people routinely spend time for working, shopping, or going out, while the last type captures the pathways between these locations. The main result of the study is the strong confirmatory evidence for crime pattern theory: every 100 visitors spending in local venues increase crime by 4.77 % and every 100 pass-through visitors en route to other locations result in an additional increase of 7.22 %. These results establish that human mobility inferred from location-based social networks (LBSNs) data can be very effective in describing crime concentrations, and advance the theoretical literature of crime, being highly relevant for criminologists and computational social scientists. Different types of routine activities vary in their relationship to crime, with leisure activities having the highest positive association with crime. Crimes are unequally impacted by the mobility flows in the area, with larcenies and vehicle thefts having the strongest positive association with the mobility in the neighborhood. Furthermore, the novel use of digital location services data proves to be an effective tool in forecasting temporal profiles of crime. In an out-of-sample setting, a machine learning model incorporating human mobility flows improves the prognostic capability of a baseline model of historical crime by 10.54 % when accounting for incoming, outgoing, and self-loop flows, and by 11.29 % when additionally accounting for pass-through flow.

Governments around the world have started to experiment with „predictive policing“, i. e. the use of predictive analytics with the aim of identifying the potential locations of criminal activity prior to such an event taking place. While the focus has hitherto been placed on areas with high population density, the last study presented in this thesis addresses the challenging undertaking of predicting crime hotspots in regions with low population densities and highly unequally-distributed crime. This results in a severe sparsity (i. e., class imbalance) of the outcome variable, which impedes predictive modeling. To alleviate this, the study proposes a machine learning approach for spatio-temporal prediction that is specifically tailored to an imbalanced distribution of the class labels. The approach consists of a hyper-ensemble model aggregating the results of several base models trained on balanced

datasets achieved through random under-sampling. With the task of predicting daily granular burglary hotspots, the model is tested in an actual setting with state-of-the-art predictors (i. e., socio-economic, geographical, temporal, meteorological, and crime variables in fine resolution), and outperforms all baselines from literature. The proposed imbalance-aware hyper-ensemble increases the hit ratio considerably from 18.1 % to 24.6 % when aiming for the top 5 % of hotspots, and from 53.1 % to 60.4 % when aiming for the top 20 % of hotspots. When predicting weekly hotspots instead of daily hotspots, the performance increases and the approach achieves a hit rate of 51.9 % at 5 % coverage level, and a hit rate of 74.4 % at 20 % coverage level. From all features sets, the locational features yield the best predictive results in the presented setup, confirming the strong relevance of social disorganization and crime pattern theories in short-term crime prediction applied to low population density areas. As direct implications, the findings help decision-makers in law enforcement and contribute to the adoption of predictive policing in low population density regions.

Zusammenfassung

Kriminalität ist allgegenwärtig und hat erheblich negative Auswirkungen sowohl auf Individuen als auch auf die Gesellschaft als Ganzes, einschliesslich finanzieller Verluste, körperlicher Schäden, psychischer Belastungen und einer allgemein verminderten Lebensqualität. Zudem tritt Kriminalität in allen Arten von Umgebungen auf: Von dicht besiedelten bis zu dünn besiedelten Gebieten, von Entwicklungsländern bis zu den am weitesten entwickelten Ländern der Welt. Daher ist es von grösster Bedeutung, unser Verständnis von Kriminalität zu verbessern und Entscheidungsträger bei der Entwicklung und Umsetzung wirksamer Massnahmen zur Kriminalitätsprävention zu unterstützen.

Verbrechen ist kein Zufall, sondern das Ergebnis einer komplexen Interaktion zwischen Tätern, Zielen und der Umwelt. Verschiedene Arten von Verbrechen, darunter sowohl Eigentumsdelikte als auch Gewaltverbrechen, folgen dabei raumzeitlichen Mustern. Langfristige Kriminalitätsmuster können über Jahre und sogar Jahrzehnte anhalten, während ein erhöhtes kurzfristiges Risiko über einen Zeitraum von Tagen bis Wochen festgestellt werden kann. Daher haben sowohl lang- als auch kurzfristige Modelle zur Kriminalitätsvorhersage einen theoretischen und praktischen Wert.

Das Aufkommen von urbanen und sozialen Big Data, gepaart mit Innovationen auf dem Gebiet des Maschinellen Lernens, hat die Forschung in den aufstrebenden Bereichen der Computational Social Science und des Urban Computing vorangetrieben. Diese Forschung hat zu einem besseren Verständnis des menschlichen Verhaltens geführt und zur Entwicklung von analytischen Lösungen für Herausforderungen in unseren Städten. Diese interdisziplinäre Arbeit erschliesst dieses Potenzial und untersucht Forschungsmöglichkeiten, die sich aus der Entwicklung oder der Anwendung moderner Techniken des Data Science auf allgegenwärtige Daten ergeben. Es werden mehrere Studien vorgestellt, die darauf abzielen, das Verständnis der raumzeitlichen Konzentration von Kriminalität in den drei Dimensionen Daten, Modelle und Anwendungen voranzutreiben. Die Vorteile für das Erforschen von Kriminalität sind zweierlei: (1) Es werden Innovationen im Bereich raumzeitlicher Daten und Methoden zur lang- und kurzfristigen Kriminalitätsvorhersage und deren Anwendungen und Auswirkungen auf Forschung und Praxis vorgestellt, und (2) es wird eine neuartige Anwendung von Daten aus ortsbezogenen Diensten vorgestellt, um eine sehr detaillierte und

skalierbare Operationalisierung der Kriminalitätstheorie zu erreichen. Dies verbessert unser allgemeines Verständnis von Kriminalität.

Mehrere Kriminalitätstheorien, einschliesslich der Social Disorganization Theory, haben vorgeschlagen, dass Kriminalitätskonzentrationen bis zu einem gewissen Grad durch die sozialen Eigenschaften der Nachbarschaft erklärt werden können (wie Einkommensniveau, Wohninstabilität und ethnische Heterogenität). Diese Theorien wurden in empirischen Studien mit Hilfe von Volkszählungsstatistiken, welche die Wohnbevölkerung der jeweiligen Nachbarschaft beschreiben, umfassend validiert. Andere Theorien, wie die Crime Pattern Theory, haben eine Verbindung zwischen Kriminalität und den Eigenschaften und Aktivitäten der lokalen Bevölkerung vorgeschlagen. Doch bisher hatten die entsprechenden empirischen Studien nur begrenzte Daten zur Verfügung. Daher wertet die erste Studie in dieser Arbeit die Integration verschiedener urbaner Datenquellen, welche die Aktivitäten der umgebenden Bevölkerung erfassen, in Modelle zur langfristigen Vorhersage von Kriminalität aus. Mit dem Ziel, jährliche Kriminalitätszahlen in einer Nachbarschaft vorherzusagen, wendet die Studie mehrere Entscheidungsbaum-basierte Ensemblemodelle aus dem Maschinellen Lernen an. Zu diesem Zweck wurde eine grosse und heterogene Menge von Merkmalen aus Foursquare Locations und Check-ins, U-Bahn-Ausgängen und Eingängen und Taxifahrten erstellt. Die Modellspezifikationen inklusive der menschlichen Aktivitätsmerkmale verbessern immer die Baselines, welche nur Zensusdaten verwenden, und dies für alle Arten von Kriminalität und für verschiedene Auswertungen. Zum Beispiel erhöht sich die R2-Kennzahl um 30 Prozentpunkte bei der Vorhersage von Kriminalität in zuvor nicht betrachteten Stadtteilen (geographisches Testset) und um 7 Prozentpunkte bei der Vorhersage von Kriminalität im folgenden Jahr (zeitliches Testset). Die Genauigkeit der Vorhersage variiert durch das Hinzufügen der neuen Merkmalen je nach Kriminalitätsart: Den grössten Zugewinn erlangt man bei schweren Diebstählen, während Übergriffe durch die Zensusdaten bereits gut vorhergesagt werden. Insbesondere liefern diejenigen Modelle, welche Merkmale aus den Foursquare-Daten verwenden, die höchste prognostische Kapazität im Vergleich zu Modellen, welche Merkmale aus den Taxi- oder U-Bahn-Daten verwenden. Die grösste Verbesserung durch die Verwendung menschlicher Aktivitätsmerkmale wurde in belebten Stadtvierteln mit zum Beispiel Freizeit- oder Einkaufszonen beobachtet. Neben der

Weiterentwicklung der Urban-Computing-Literatur bieten diese Ergebnisse wertvolle Erkenntnisse für die Verantwortlichen der Stadtentwicklung.

Frühere Studien, einschliesslich der ersten Studie dieser Arbeit, beschränken sich auf aggregierte Statistiken der Besucher bei der Beurteilung der Umgebungspopulation in einem Gebiet. Aus diesem Grund vernachlässigen sie die zeitliche Dynamik der einzelnen menschlichen Bewegungen. Diesen Punkt adressiert die zweite Studie dieser Arbeit. Sie präsentiert die erste Untersuchung, in wie weit granuläre Daten menschlicher Mobilität die raumzeitliche Konzentration von Kriminalität beschreiben und vorhersagen können. Zu diesem Zweck schlägt die Studie die Verwendung von Daten aus standortbezogenen Diensten, insbesondere Foursquare, vor. Diese Art von Daten besteht aus einzelnen Bewegungen und bietet die Möglichkeit, zwischen verschiedenen Arten von Mobilitätsströmen zu unterscheiden: (1) ein- oder ausgehend aus einer Nachbarschaft, (2) in ihr verbleibend, oder (3) Bewegungen, bei denen Menschen sich durch eine gegebene Nachbarschaft hindurchbewegen. Dementsprechend modellieren die ersten beiden Typen von Mobilitätsströmen das Konzept der routinemässigen Aktivitätsknoten aus der Crime Pattern Theory, d.h. Orte, an denen Menschen routinemässig Zeit zum Arbeiten, Einkaufen oder Ausgehen verbringen, während der letzte Typ die Wege zwischen diesen Orten erfasst. Das Hauptergebnis der Studie ist die starke Bestätigung der Kriminalitätstheorie: Pro 100 Besucher, die zusätzlich an einem Standort verbleiben, erhöht sich die Kriminalität um 4.77 % und pro 100 Durchreisenden auf dem Weg zu anderen Standorten erhöht sich die Kriminalität um 7.22 %. Diese Ergebnisse zeigen, dass die menschliche Mobilität, die aus den LBSN-Daten abgeleitet wird, sehr effektiv bei der Beschreibung der Kriminalkonzentration sein kann und die theoretische Literatur der Kriminalität vorantreibt. Dies ist für Kriminologen und Informatiker von hoher Relevanz. Verschiedene Arten von Routinetätigkeiten unterscheiden sich in ihrem Verhältnis zur Kriminalität, wobei Freizeitaktivitäten den höchsten positiven Zusammenhang mit Kriminalität aufweisen. Verbrechen werden ungleichmässig von den Mobilitätsströmen in der Region beeinflusst, wobei Diebstähle und Fahrzeugdiebstähle den stärksten positiven Zusammenhang mit der Mobilität in der Nachbarschaft haben. Darüber hinaus erweist sich die neuartige Nutzung digitaler Standortdaten als wirksames Instrument zur Vorhersage zeitlicher Kriminalitätsprofile. In einer Out-of-Sample Evaluation verbessert ein ma-

schinell lernendes Modell unter Verwendung menschlicher Mobilitätsströme die Prognosefähigkeit eines Baseline Modells der historischen Kriminalität um 10.54 % bei der Berücksichtigung von Eingangs- oder Ausgangs- und Verbleibsströmen und um 11.29 % bei der zusätzlichen Berücksichtigung von Transitströmen.

Regierungen auf der ganzen Welt haben begonnen, mit „Predictive Policing“ zu experimentieren, das heisst dem Einsatz von analytischen Methoden zur Identifikation von Orten krimineller Aktivität, bevor eine solche überhaupt stattgefunden hat. Während der Schwerpunkt hier bisher auf Gebiete mit hoher Bevölkerungsdichte gelegt wurde, befasst sich die letzte in dieser Arbeit vorgestellte Studie mit dem anspruchsvollen Unterfangen, Kriminalitäts-Hotspots in Regionen mit geringer Bevölkerungsdichte und sehr ungleichmässig verteilter Kriminalität vorherzusagen. Dies führt zu einer starken Sparsity (Klassenungleichgewicht) der Ergebnisvariablen, was die prädiktive Modellierung erschwert. Um dieses Problem zu adressieren, schlägt die Studie einen Ansatz maschinellen Lernens für die raumzeitliche Vorhersage vor, der speziell auf eine unausgewogene Verteilung der Klassenlabel zugeschnitten ist. Der Ansatz besteht aus einem Hyper-Ensemble-Modell, das die Ergebnisse mehrerer Basismodelle aggregiert, welche auf ausgewogenen Datensätzen, die durch zufälliges Under-Sampling kreiert wurden, trainiert worden sind. Mit der Aufgabe, tägliche granuläre Hotspots von Einbrüchen vorherzusagen, wird das Modell in einem Setting mit modernsten Prädiktoren (sozioökonomischen, geografischen, zeitlichen, meteorologischen und kriminellen Variablen in feiner Auflösung) getestet und übertrifft alle Baselines aus der Literatur. Das hier vorgeschlagene Hyper-Ensemble erhöht die Trefferquote deutlich von 18.1 % auf 24.6 %, wenn es um die Top 5 % der Hotspots geht, und von 53.1 % auf 60.4 %, wenn es um die Top 20 % der Hotspots geht. Bei der Vorhersage von wöchentlichen Hotspots anstelle von täglichen Hotspots steigt die Performance und der Ansatz erreicht eine Trefferquote von 51.9 % bei 5 % Abdeckung und eine Trefferquote von 74.4 % bei 20 % Abdeckung. Von allen Merkmalen liefern die Standortmerkmale die besten prädiktiven Ergebnisse in der dargestellten Konfiguration und bestätigen die starke Relevanz der Social Disorganization Theory und der Crime Pattern Theory für die kurzfristige Kriminalitätsvorhersage in Gebieten mit geringer Bevölkerungsdichte. Als direkte Auswirkung helfen die Ergebnisse den Entscheidungsträgern bei der Strafverfolgung und tra-

gen zur Einführung der prädiktiven Polizeiarbeit in Regionen mit geringer Bevölkerungsdichte bei.