Bode, Felix
Stoffel, Florian
Keim, Daniel

April 2017

Variabilität und Validität von Qualitätsmetriken im Bereich von Predictive Policing

Diskussionen zu prädiktiven Kriminalitätsanalysen, auch unter dem Begriff Predictive Policing bekannt, prägen aktuell die kriminologische und polizeiwissenschaftliche Literatur. Hersteller entsprechender Softwarelösungen berichten von vielversprechenden Ergebnissen. Medienberichte unterschiedlichster Art setzen sich mit der operativen Umsetzung – mehr oder weniger kritisch und vielfältig – auseinander. In der Folge werden auf politischer Ebene vermehrt Entscheidungen getroffen, Predictive Policing in die Sicherheitsorgane des Bundes und der Länder zu implementieren. Das Spektrum der Umsetzungsmöglichkeiten reicht dabei von pragmatischen Einzellösungen, bis hin zu stark theoretisch basierten, wissenschaftlichen Vorgehensweisen. Fast allen Umsetzungen und damit verbundenen Veröffentlichungen zur Wirksamkeit ist gemein, dass sich die Qualitätsmetriken auf Trefferraten (Hit Rates) beziehen. Sie prüfen auf unterschiedliche Art und Weise, ob in einem vorher prognostizierten Bereich tatsächlich ein entsprechendes Delikt passiert ist. Doch: Was steckt hinter diesen Zahlen? Wie lässt sich welche Trefferrate deuten und ist ein Vergleich mit Trefferraten anderer Predictive-Policing-Umsetzungen überhaupt möglich?

Mit diesem Beitrag sollen die Variabilität und Validität von Qualitätsmetriken im Bereich von Predictive Policing dargestellt und kritisch hinterfragt werden. Neben einer einleitenden Darstellung des methodischen Predictive-Policing-Prozesses werden die wesentlichen vorhandenen Möglichkeiten zur Trefferratenberechnung systematisch dokumentiert und visuell dargestellt. Anschließend erfolgt eine kritische Auseinandersetzung, ob und in welchem Ausmaß entsprechende Variabilität und Validität vorhanden sind.

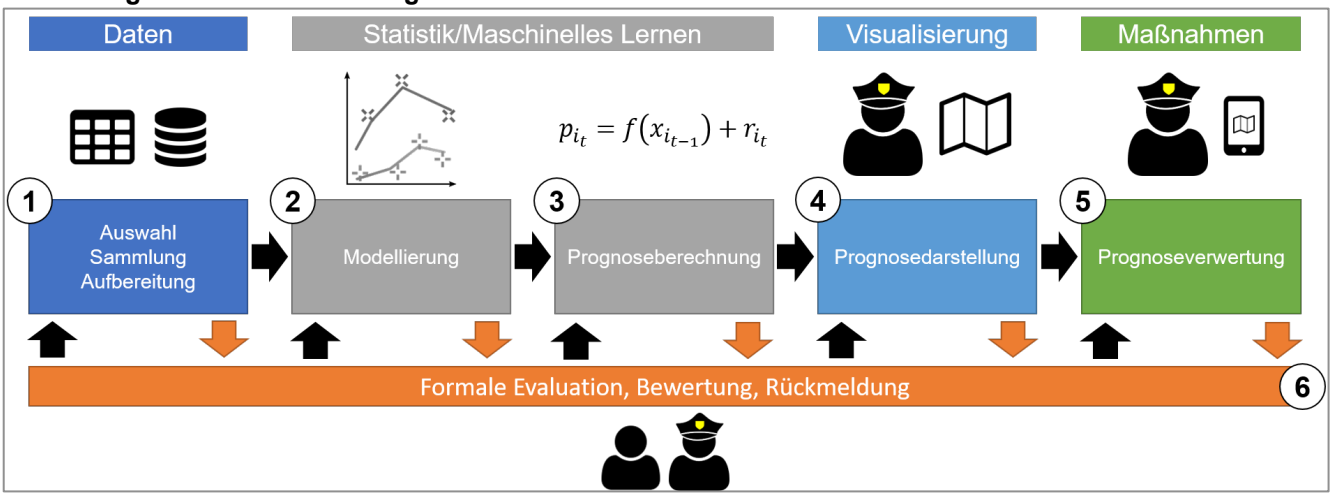
In der kriminologischen und polizeiwissenschaftlichen Literatur besteht grundsätzlich Einigkeit darüber, was begrifflich unter Predictive Policing zu verstehen ist, da es sich aus dem Englischen „to predict“ (vorhersagen) und „policing“ (Polizeiarbeit) ableitet. Gegenstand der Vorhersagen sind unterschiedliche Kriminalitätsphänomene (vgl. Berk et al. 2009; Saunders et al. 2016). In diesem Aufsatz wird die Prognostizierung von möglichen Räumen mit erhöhtem Risiko für bestimmte Deliktsbereiche fokussiert.

Der Predictive-Policing-Prozess

Der Prozess, der sich um Predictive Policing konstituiert, wird häufig mit dem von Perry et al. (2013: 12) entwickelten „Prediction-Led Policing Business Process“ dargestellt. Dieser betrachtet und dokumentiert den Kreislauf und die Wechselwirkungen, welche mit Umsetzungsmöglichkeiten von Predictive Policing verbunden sind. Die Ausgestaltung eines methodischen Prozesses findet aber in dieser Darstellung keine Anwendung, ebenfalls findet sich hierzu in der weiterführenden Fachliteratur nichts. Um dieses methodische Defizit zu schließen, lässt sich der methodische Prozess von Predictive Policing wie in Abbildung 1 gezeigt illustrieren[1]. Aufgeteilt in sechs Schritte, bietet dieser Prozess eine Einsicht in die einzelnen Prozesse zur Umsetzung von Predictive Policing aus polizeilicher Sicht. Abweichungen davon sind natürlich denkbar, jedoch dürften zumindest ähnliche Ausgestaltungen immer dann vorliegen, wenn Techniken des maschinellen Lernens angewendet werden. Ebenso wird besonderes Augenmerk auf eine kontinuierliche Evaluation bzw. eine durchgängige Möglichkeit für Rückmeldungen gelegt.

[1] Zugunsten einer besseren Lesbarkeit und um einfacher auf die verschiedenen Schritte im Prozess zurückgreifen zu können, werden nachfolgend die Zahlen als Ziffern dargestellt und nicht ausgeschrieben.

Abbildung 1: Predictive-Policing Prozess



Schritt 1: Daten

Der Prozess beginnt im ersten Schritt mit der Sichtung und Auswahl von Datenquellen sowie der Sammlung und Aufbereitung von Datensätzen, die in den weiteren Bestandteilen des Prozesses verarbeitet werden. Mit entsprechender Software ist es möglich, verschiedene Datenquellen miteinander in Beziehung zu setzen. Zentral sind hierbei die raum- und zeitbezogene Zusammenführung. Dabei können polizeiliche Vorgangsdaten mit nicht-polizeilichen Daten (z. B. Daten zur Wetterlage, Wohnlage oder Entfernung zur nächstgelegenen Autobahn) kombiniert werden. In diesem Zusammenhang ist wichtig, dass alle ausgewählten Daten geografisch referenziert werden können, sodass ein einheitlicher, maschinell verarbeitbarer Datensatz vorliegt, der die Basis für Predictive Policing darstellt.

Zur weiteren Umsetzung existieren eine große Menge unterschiedlicher Konzepte und Methoden, angefangen bei rein auf den Near-Repeat-Ansatz basierten Lösungen (vgl. Balogh 2016) bis hin zu Ansätzen, die sich stärker auf wissenschaftliche Theorien stützen (vgl. Pollich & Bode 2017). Trotz der mannigfachen methodischen Herangehensweisen haben die unterschiedlichen Konzepte derzeit viele Gemeinsamkeiten: Es werden in der Regel keine personenbezogenen Daten benutzt, die Fokussierung auf den Wohnungseinbruchdiebstahl ist zentral und die Berechnungsmodelle nutzen für die Prognoseberechnung im Vorfeld historische Daten.

Die beiden Folgeschritte lassen sich unter den beiden Begriffen „Statistik“ bzw. „Maschinelles Lernen“ subsumieren. Sie beinhalten die konkrete methodische Ausgestaltung des Predictive Policings. Diese wird, wie in der automatischen Datenanalyse und Prognose üblich, stets in zwei Bestandteile zerlegt und deshalb auch in der hier erstellten Prozessvisualisierung in Schritt 2 und 3 aufgeteilt.

Schritt 2: Modellierung

Zu Beginn wird ein konkretes Modell unter Verwendung der vorliegenden historischen Daten erstellt, um die Kriminalitätslage möglichst angemessen in allen gewünschten Facetten abzubilden. Beispielsweise wäre eine Modellierung mittels Regressionen (vgl. Box et al. 2015: 305 ff.), Entscheidungsbäumen (z. B. Chi-square Automated Interaction Detection [CHAID], vgl. Kass 1980) oder künstlicher neuronaler Netze möglich (vgl. Zhang & Qi 2005).

Schritt 3: Prognoseberechnung

Das erstellte Modell wird auf aktuelle bzw. mögliche zukünftige Daten angewandt, um die Wahrscheinlichkeit eines bestimmten Delikts in einer geografischen Bezugsgröße zu ermitteln. Dieser Schritt stellt die eigentliche Prognoseberechnung dar und ist, mit den aus Schritt 2 gewonnen Erkenntnissen, das Herzstück im Predictive-Policing-Prozess. Das Ergebnis der Berechnung präsentiert sich regelmäßig in einer Auswahl an geografischen Räumen, die ein höheres Kriminalitätsrisiko aufweisen als andere Räume im gleichen Prognosezeitraum. Da sich dieses erhöhte Kriminalitätsrisiko stets auf einen bestimmten Prognosezeitraum und auf bestimmte Prognosegebiete bezieht, lassen sich solche Wahrscheinlichkeiten auch als raum-zeitliche Prädispositionsfaktoren bezeichnen.

Schritt 4/5: Prognosedarstellung und -verwertung

Schritt 4 sieht die adäquate Darstellung der Kriminalitätsprognosen vor, um sie sodann im Feld (meist durch operative Polizeieinheiten) einzu-

setzen (Schritt 5). Mit zunehmender Digitalisierung sind neben Karten im Papierformat zwischenzeitlich auch Prognosevisualisierungen auf Tablet-PCs oder Mobilfunkgeräten möglich. Das Forschungsfeld der Visual Analytics (vgl. Keim et al. 2010) kann hier besondere Flexibilität und Aufgabenangemessenheit garantieren.

Schritt 6: Evaluation
Schritt 6 umfasst, bezogen auf den methodischen Predictive-Policing-Prozess, die durchgehende Evaluation und Bewertung der angewandten Methoden. Es handelt sich um die formale, statistische Beschreibung von beobachteten Effekten, z. B. mit der Berechnung von Trefferraten als auch durchgehende Plausibilitätsprüfungen und Ad-Hoc Verifikationen von Zwischenergebnissen, um beispielsweise die Eignung der gewählten Methode in den Schritten 2 und 3 oder der gewählten Visualisierungstechnik in Schritt 4 kontinuierlich sicherzustellen. Da solche Rückmeldungen grundsätzlich für die Schritte 1 bis 5 möglich und sinnvoll sind, ist dies mit durchgehenden Eingaben (orangene Pfeile) und Ausgaben (schwarze Pfeile) in Schritt 6 visualisiert.

Einfluss von Datenqualität auf den Predictive-Policing-Prozess

Der Predictive-Policing-Prozess besteht aus mehreren aufeinander aufbauenden Einzelschritten (Abbildung 1). Alle Prozessbestandteile hängen von den zu verarbeitenden Daten, deren Sammlung sowie der Aufbereitung zur maschinellen Weiterverarbeitung ab (Schritt 1).

Geschehen in Schritt 1 Fehler, sog. *Datenfehler*, wirken sich diese mit unterschiedlichen, später nur schwer zu rekonstruierenden Folgen auf den gesamten Predictive-Policing-Prozess aus. Es lässt sich leicht nachvollziehen, dass dadurch im Fehlerfall die Verlässlichkeit des gesamten Prozesses und damit der Ergebnisse in Frage steht. Gleichzeitig ist damit die Validität der im Folgeabschnitt noch darzustellenden Qualitätsmetriken unklar. Neben inhaltlichen Fehlerquellen, z. B. der inadäquaten Auswahl von Datensätzen, beispielsweise Datenquellen die in keinem Kausalzusammenhang mit dem vorherzusagenden Delikt stehen, sind ebenso Probleme auf technischer Ebene denkbar. Dies können fehlerhafte Datenimportroutinen aufgrund unzureichend spezifizierter Datenformate sein als auch die Wahl eines ungeeigneten Verfahrens zur Zusammenführung.

Abgesehen von diesen konkreten Problemen, die in Schritt 1 des Prozessmodells auftreten können, spielt bei der Bewertung von Predictive Policing ebenso die Datenqualität im Hinblick auf den Begriff der *Datenunsicherheit* (vgl. Morgan et al. 1990; Fritsch et al. 1998; Kinkeldey et al. 2014) eine große Rolle. Der Begriff der Unsicherheit bezieht sich dabei auf das Problem, dass meist nicht bekannt ist, in welchem Umfang Fehler in den verwendeten Daten enthalten sind. Dies kann sich verschieden äußern. Denkbar sind in diesem Zusammenhang Erfassungsprobleme (Messunsicherheiten), beispielsweise bei der Erfassung der Tatzeit eines Wohnungseinbruchs oder ein schwankender Präzisionsgrad, wie er bei der satellitengestützten Bestimmung von Geokoordinaten vorkommen kann. Auf polizeilicher Seite kann sich dies darin äußern, dass Straftaten von Geschädigten bei der Aufnahme zunächst rechtlich falsch bewertet oder aber von den Geschädigten verspätet zur Anzeige gebracht werden. Die auf Grundlage dieser Datenbasis angewandten Qualitätsmetriken müssen insofern revidiert werden. Da gerade das Delikt Wohnungseinbruch diesen Unsicherheiten unterliegt, ist die Berechnungsgrundlage für sämtliche anzuwendenden Qualitätsmetriken schon im Vorfeld hoch variabel. So ist es beispielsweise nicht unüblich, dass Einbrüche von Geschädigten verspätet (z. B. nach einem Urlaub) oder gar nicht zur Anzeige gebracht werden.

Nach der Datenerfassung sind solche Fehlerquellen nur schwer bis unmöglich zu rekonstruieren und zu korrigieren. Basiert die Kriminalitätsprognose auf sehr unsicheren bzw. fehlerhaft erfassten Daten, steht auch dessen objektive Bewertung grundsätzlich in Frage.

Ein generelles Problem von Predictive Policing mittels automatischen Datenanalysemethoden betrifft zudem deren grundlegende Annahme, dass das Delikt, das Gegenstand der Analyse ist, mit den vorliegenden Daten hinreichend genau im Hinblick auf Einflussfaktoren wie Raum, Zeit oder lokale Gegebenheiten beschrieben ist. Dies ist Voraussetzung für deren Anwendung, denn Verfahren zur automatischen Datenanalyse bzw. des maschinellen Lernens setzen die ausreichende Messbarkeit des analysierten Phänomens voraus. Ohne diese Vorbedingung wären objektive Entscheidungsgrundlagen nicht algorithmisch formulierbar, was zur Berechnung von validen Ergebnissen aber notwendig ist. Meist wird in die generierten Entscheidungsgrundlagen ein Residuum oder eine *Restgröße* einbezogen, die jene Teile des Phänomens abdeckt, die nicht

ausreichend aus den zur Verfügung stehenden Daten erklärbar oder Gegenstand eines nicht erklärbaren Prozesses (Zufallsprozess oder nicht erfassten Einflussfaktoren) sind. Die aktuell beobachtbaren guten Ergebnisse der automatischen Datenanalyse bzw. des maschinellen Lernens in vielen Anwendungsfeldern sind meist darin begründet, dass die in der Modellierung enthaltenen Residuen recht klein sind. Dies ist insbesondere dann der Fall, wenn große Datenmengen analysiert werden, die ein Phänomen in möglichst vielen Ausprägungen und Varianten beschreiben können. Im polizeilichen Umfeld ist dieser Umstand jedoch meist nicht gegeben, da die Vorgangsbearbeitungssysteme sehr unterschiedlich sind und stets menschliches Verhalten in Raum und Zeit abgebildet wird. Ebenso ist die vollständige objektive Beschreibung von Kriminalitätsphänomenen nicht umfassend möglich, insbesondere dann, wenn nicht beobachtbare oder nicht quantifizierbare Effekte, z. B. aus dem nicht öffentlichen Umfeld potenzieller Täter oder Tatgelegenheiten, eine Rolle spielen. In diesem Zusammenhang sind Tatzeiterfassungen denkbar, denn nicht immer lässt sich bestimmen, wann ein bestimmtes Delikt genau passiert ist. Dieses grundlegende Problem, das sich stringent durch sämtliche Predictive-Policing-Umsetzungen zieht, wirft die Frage auf, ob angewendete Methoden zur Bewertung der Qualität von Predictive Policing ohne Angabe von Fehlerraten bzw. der ungefähren Größenordnung des Residuums, überhaupt als valide gelten können. Dies soll nachfolgend mit der Darstellung von geografischen Bezugsgrößen als Prognoseräume und der anschließenden Dokumentation und Diskussion von Qualitätsmetriken näher betrachtet werden.

Prognosegebiete als geografische Bezugsgröße

Die Erstellung von raum-zeitlichen Kriminalitätsprognosen muss sich, wie zuvor bereits angedeutet, immer auf einen bestimmten georeferenzierten Raum, den sog. Prognoseraum, beziehen (Schritt 2 und 3). Eine Prognoseerstellung lediglich georeferenziert auf das einzelne Delikt und dessen Verortung im Raum ist zwar statistisch möglich, birgt aber keinen Erkenntnis- und Prognosegewinn. Denn der Eintritt eines kriminellen Ereignisses (z. B. ein Einbruch an einer bestimmten Adresse) ist im Verhältnis zu allen anderen in Frage kommenden Räumen (beispielsweise alle anderen potenziellen Adressen in einer Stadt, an denen ein Einbruch auftreten könnte) zu selten. Anders ausgedrückt: Die statistische Wahrscheinlichkeit, dass am nächsten Tag in Haus X eingebrochen wird, tendiert gegen 0. Die statistische Wahrscheinlichkeit, dass im Stadtviertel Y, in dem auch Haus X liegt, am nächsten Tag eingebrochen wird, ist dagegen deutlich höher. Die Erstellung von Kriminalitätsprognosen muss sich folglich immer auf einen Raum beziehen, der sich bildlich auf die tatsächlichen kriminellen Ereignisse legt. Auf diese geografische Bezugsgröße werden alle in der Vergangenheit dokumentierten, darin liegenden Kriminalitätsereignisse aggregiert.

Die Wahl einer geeigneten geografischen Bezugsgröße als Prognoseraum richtet sich nach dem zu prognostizierenden Delikt und nach den Bedürfnissen des jeweiligen Endnutzers. Ein Prognoseraum sollte immer so groß gewählt werden, dass operative Polizeikräfte diesen auch gut überwachen können. Dies wird in der Praxis bei bestehenden Prognosemodellen und entsprechenden Softwarelösungen meist mittels Gitterzellen bestimmter Größen (z. B. 250m x 250m) oder homogen abgegrenzten Räumen (z. B. Wohnquartieren[2]) gelöst. Die Praxis mit Gitterzellen zu arbeiten ist die bei Predictive Policing wohl prominenteste Variante, den Raum einer Stadt in Prognosegebiete aufzuteilen, Kriminalitätsereignisse zuzuordnen und raum-zeitliche Prädispositionsfaktoren für unterschiedlichste Arten von Kriminalität zu berechnen. Alternativ bietet es sich an, zur räumlichen Einteilung Wohnquartiere zu verwenden. Diese haben den Vorteil, dass sie die Stadt mit ihren natürlichen, im Laufe der Zeit gewachsenen Grenzen nicht willkürlich durchschneiden, sondern sich an sozialen Räumen und geografischen Barrieren orientieren[3]. Exemplarisch wird dies in Abbildung 2 und Abbildung 3 deutlich. Die Wohnquartiere (Abbildung 3) wurden für diesen Zweck mit dem Clusteralgorithmus DBSCAN (Density-Based Spatial Clustering of Applications with Noise, vgl. Ester et al. 1996) berechnet[4].

[2] Unter Wohnquartier wird eine bestimmte Anzahl an Haushalten mit großer Homogenität subsumiert. Die Geo Marketing GmbH Nexiga versteht hierunter durchschnittlich 400 Haushalte mit größtmöglicher Homogenität, deren Ursprung in den Wahlbezirken liegt (vgl. Nexiga 2017). Je nach Anzahl der Haushalte kann es im innerstädtischen Bereich zu geografisch kleineren, im städtischen Randbereich zu geografisch größeren Wohnquartieren kommen.

[3] An dieser Stelle sei auf das Modifiable Areal Unit Problem (MAUP) verwiesen, das die Problematik verschiedener geografischer Raumpartitionen auf statistische Modellierung und Datenanalyse beschreibt (vgl. Openshaw 1984).

[4] In diesem Zusammenhang werden einzelne Elemente eines Datensatzes, basierend auf ihren Entfernungen, die an den Anwendungsfall angepasst berechnet werden, automatisch in Gruppen eingeteilt.

Abbildung 2: Illustration Gitterzellen (250m x 250m)

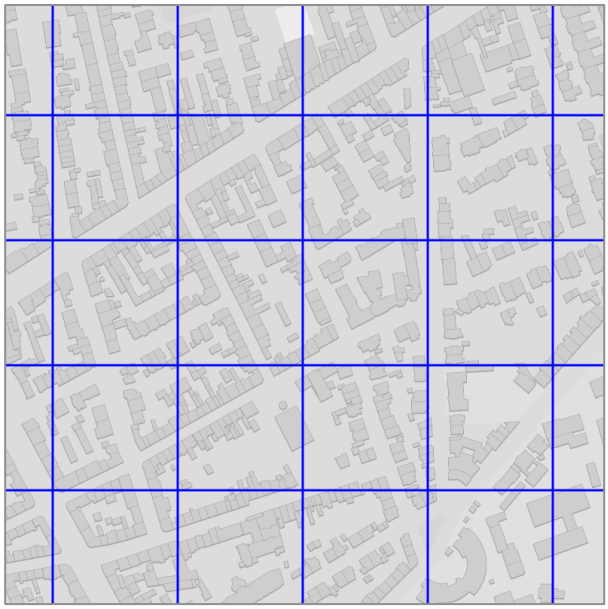


Abbildung 3: Illustration Wohnquartiere

Klassische Qualitätsmetriken

Mit der Erstellung von Kriminalitätsprognosen kommt unabhängig von dem verwendeten kriminologischen und mathematischen Modell die Frage auf, ob das erwartete Ereignis eingetreten ist. Seit der Anwendung prädiktiver Kriminalitätsanalysen beschäftigt diese Frage Anwender wie Kritiker, die stets auf der Suche nach einem Qualitätsmaß sind, welches diese Anforderung objektiv, valide und zuverlässig erfüllt. Dies gestaltet sich jedoch schwierig, da bei Erstellung von Prognosen grundsätzlich mit Wahrscheinlichkeiten operiert wird. Diesen ist immer ein gewisses Maß an Unsicherheit inhärent. Bei statistisch basierten Predictive-Policing-Verfahren werden Merkmale aus dem historischen Datensatz auf mögliche Korrelationen mit dem zu prognostizierenden Delikt unter Beachtung der Stärke des Zusammenhangs untersucht. Es wird versucht, beispielsweise für das Delikt Wohnungseinbruch, typische Prädiktoren zu identifizieren. Liegen diese in einem bestimmten Raum vor, kann dort von einer erhöhten Kriminalitätswahrscheinlichkeit ausgegangen werden. Beispiele für solche Merkmale sind die Tageszeit, der Wochentag oder der Modus Operandi. Die festgestellten Zusammenhänge sind damit eine modellhafte Abbildung der Datenbasis aus der Vergangenheit, welche auf die Zukunft (Prognosezeitraum) übertragen wird. Kommt das zu prognostizierende Delikt selten vor, äußert sich dies in einer generell geringen Eintrittswahrscheinlichkeit; man spricht von „statistisch seltenen Ereignissen“. Selbst beim Vorliegen von statistisch bedeutsamen Prädiktoren und damit einer Erhöhung der Eintrittswahrscheinlichkeit eines Delikts in einem gewissen Gebiet, bleibt die Wahrscheinlichkeit, dass dieses eintritt, naturgemäß klein. Geht man z. B. davon aus, dass durchschnittlich etwa 5% aller Wohngebiete von Einbrüchen betroffen sind, könnte eine Steigerung auf 25% in bestimmten Gebieten das Resultat eines Prognosemodells sein. Das würde einer Verfünffachung des Ausgangsrisikos entsprechen. Dennoch wird mit einer deutlich höheren Wahrscheinlichkeit, nämlich zu 75%, in diesen Gebieten kein Wohnungseinbruch begangen. Das erklärt die (in der Regel relativ häufige) Beobachtung, dass das prognostizierte Delikt im ausgesuchten Prognosezeitraum nicht eintritt. Die Messung mit einer entsprechenden Qualitätsmetrik versucht jedoch genau diese (immer noch unwahrscheinlichen) Ereignisse zu erfassen.

Weiterhin besteht die Besonderheit, dass im Zuge der Prävention gegen den Eintritt des prognostizierten Delikts gearbeitet wird. Das bedeutet, die Polizei arbeitet stets daran, dass das prognostizierte Ereignis nicht eintritt. Dies muss bei der Evaluation, Anwendung und Interpretation entsprechender Qualitätsmetriken stets berücksichtigt werden.

Nachfolgend werden die gängigen, in der kriminologischen und polizeiwissenschaftlichen Literatur immer aufkommenden Berechnungsmodelle für Predictive Policing dargestellt[5]. Exemplifiziert wird diese Darstellung mit dem Delikt Wohnungseinbruchdiebstahl (WED), da dieses in der Diskussion zu Predictive Policing derzeit fokus-

[5] Die Darstellung ist nicht abschließend.

siert wird. Darüber hinaus bietet es sich an, den WED auch deshalb auszuwählen, da er – durch die Annahme, dass menschliches Verhalten gewissen Mustern folgt – dynamisch in Raum und Zeit innerhalb des gesamten Stadtgebietes passiert. Als Kontrast: Würde versucht werden, Taschendiebstahl in einer Großstadt zu prognostizieren, würden die Analysen sicherlich dazu führen, dass die Prognosen immer wieder auf die gleichen, öffentlich belebten Orte gelegt würden und somit wenig neues Wissen im Sinne prädiktiver Analysen produzieren. Gleiches gilt für Kriminalität im Bereich der sog. Kontrolldelikte, wie z. B. Betäubungsmittelkriminalität. Prognosen würden sich stets auf Gebiete beziehen, in denen die Polizei zuvor aktiv kontrolliert hat.

Bei den hier verwendeten Kriminalitätsprognosen handelt es sich um fiktive Beispiele, um die Qualitätsmetriken entsprechend zu veranschaulichen. Die Fallzahlen für WED sind real und dem Wohnungseinbruchradar aus NRW entnommen (vgl. exemplarisch Polizei Köln 2017). Die Berechnung erfolgte anhand von Wohnquartieren.

Die Trefferrate (Hit Rate)

Die (absolute) Hit Rate (HR), auch Accuracy, ist die wohl am häufigsten dokumentierte und aufgegriffene Trefferrate im Bereich von Predictive Policing (vgl. beispielsweise Berk 2008; Chainey et al. 2008; Hunt et al. 2014; Mohler et al. 2015). Sie bezieht sich immer auf einen bestimmten Zeitraum (t), wie einen Monat, eine Woche, oder einen Tag und berechnet sich wie folgt:

$$\text{HR}[\%] = \frac{\text{Anzahl Delikte in den Prognosegebieten}}{\text{Anzahl aller Delikte}} \cdot 100$$

Sind beispielsweise in einem Prognosezeitraum von einer Woche in einer Stadt 100 WED aufgetreten und lagen 15 WED in vorher herausgegebenen Prognosegebieten, so ergibt sich mit der Berechnung $15/100 \cdot 100$ eine (absolute) Hit Rate von 15%. Anders ausgedrückt wurden 15% aller in der vergangenen Woche aufgetretenen WED richtig vorhergesagt. Die Berechnung ist einfach und sie lässt sich leicht nachvollziehen. Mit nur wenigen Daten zum Kriminalitätsgeschehen und zum Prognoseraum ist eine Messung möglich. Die Bezugsgröße, bisher die Anzahl aller Delikte, kann aber auch angepasst werden. Eine Variante ist die Anpassung der Hit Rate auf die Anzahl der Prognosegebiete (HR_P) und berechnet sich wie folgt:

$$\text{HR}_\text{P}[\%] = \frac{\text{Anzahl getroffener Progonosegebiete}}{\text{Anzahl herausgegebener Prognosegebiete}} \cdot 100$$

Im Beispiel mit den 15 WED in den Prognosegebieten würde sich, wenn insgesamt 20 Prognosegebiete zuvor herausgegeben worden wären, eine Hit Rate auf die herausgegebenen Prognosegebiete von 75% ergeben, wenn jeweils ein Delikt pro Prognosegebiet auftritt (Mehrfachtreffer dürfen nicht gezählt werden). Auf die Formel angewandt: $15 / 20 \cdot 100$. Darauf stützend lässt sich folglich die Aussage treffen, dass von allen herausgegebenen Prognosegebieten, in 75% tatsächlich ein Ereignis aufgetreten ist.

Der Predictive Accuracy Index (PAI)

Um die Hit Rate in Relation zur Größe der herausgegebenen Prognosegebiete und der Gesamtfläche der Stadt setzen zu können, haben Chainey et al. (2008) – ursprünglich für Hotspot Mapping (Brennpunktkartierung[6]) gedacht – den Predictive Accuracy Index (PAI) entwickelt. Hierbei wird die zuvor berechnete Hit Rate durch den Anteil der herausgegebenen Prognosegebiete an der Gesamtfläche der Stadt geteilt:

$$\text{PAI} = \frac{\frac{n}{N} \cdot 100}{\frac{a}{A} \cdot 100}$$

n:	Anzahl Delikte in Hotspots
N:	Anzahl aller Delikte
a:	Gesamtfläche der Hotspots
A	Gesamtfläche des betrachteten Gebiets

(vgl. Chainey et al. 2008: 14).

Am hier dokumentierten Beispiel und mit der Annahme, dass die zuvor erwähnten 20 herausgegebenen Prognosegebiete eine Fläche von 10 km² haben und die Gesamtfläche der Stadt 120 km² hat, würde sich der PAI wie folgt berechnen:

$$\frac{\text{HR}[\%]}{\text{Anteil der herausgegebener Prognosegebieten an der Gesamtfläche}} =$$

$$\frac{15}{{}^{10}/_{120} \cdot 100} = \frac{15}{8{,}3} = 1{,}81$$

Mit dem errechneten PAI-Indexwert lässt sich die Hit Rate relativiert zum Prognoseraum darstellen. Im angloamerikanischen und angelsächsischen Raum ist die Messung mit dem PAI weit verbreitet (vgl. z. B. Drawve 2014; Levine 2008, Mohler et al. 2015; Wang et al. 2012). Mit Bezug auf Hotspot Mapping wird im Rahmen der Diskussio-

[6] Brennpunktkarten basieren üblicherweise auf Techniken der Kerndichteschätzung (Silverman 1986), bei denen verschiedene Interpolationsverfahren auf ein regelmäßiges Gitter, dessen Zellen typischerweise die Häufigkeiten der betrachteten Delikte beinhalten, zur Darstellung von Hotspots angewendet werden (vgl. Braga 2005; Chainey et al. 2008).

nen zum PAI von Van Patten et al. (2009) sowie von Hart und Zandbergen (2012) die zusätzliche Nutzung des Recapture Rate Index (RRI)[7] empfohlen. Da sich Hotspot Mapping aber methodisch und inhaltlich von Predictive Policing unterscheidet (obgleich diese Begriffe in der Literatur und im allgemeinen Sprachgebrauch häufig vermengt werden), wird auf solche Qualitätsmetriken für Hotspot Mapping (RRI) hier nicht weiter eingegangen.

Der Standarized Accuracy Efficiency Index

Da mit der Bildung der Hit Rate wie auch des PAI die Güte bzw. Effektivität der herausgegebenen Prognosegebiete nicht berücksichtigt wird, wurde von Public Engines[8] in ihrem White Paper zu „Predictive Analytics vs. Hot Spotting" der Standarized Accuracy Efficiency Index (SAEI) gebildet (vgl. Public Engines 2014; Motorola Solutions 2015). Er berechnet sich aus der *Achievable Efficiency*, dem Anteil aller Delikte an allen möglichen Prognosegebieten, der *Observed Efficiency*, dem Anteil der Delikte in Prognosegebieten an allen möglichen Prognosegebieten und der *Accuracy*, die der (absoluten) Hit Rate entspricht. Konkret:

$$\text{SAEI} = \frac{(\text{Achievable Efficiency} - (\text{Observed Efficiency} \cdot \text{Accuracy}))}{\text{Achievable Efficiency}}$$

$$\text{Achievable Efficiency} = \frac{\text{Anzahl aller Delikte}}{\text{Anzahl aller möglichen Prognosegebiete}}$$

$$\text{Observed Efficiency} = \frac{\text{Anzahl Delikte in den Prognosegebieten}}{\text{Anzahl aller möglichen Prognosegebiete}}$$

$$\text{Accuracy} = \frac{\text{Anzahl Delikte in den Prognosegebieten}}{\text{Anzahl aller Delikte}} \triangleq \text{HR}$$

(vgl. Public Engines 2014: 7)

Ebenfalls am Beispiel des WED und mit der Annahme, dass in der Stadt 500 Prognoseräume vorhanden sind, gilt es, neben der bereits zuvor (absoluten) Hit Rate mit 0,15 (hier Accuracy) noch die Achievable und Observed Efficiency zu berechnen. Bei 500 Prognoseräumen wird für die Achievable Efficiency die Anzahl aller Delikte im Vorhersagezeitraum durch die Anzahl aller möglichen Prognoseräume geteilt, konkret: 100 / 500 = 0,2. Für die Observed Efficiency wird die Anzahl der Delikte in den Prognosegebieten im Vorhersagezeitraum ebenfalls durch die Anzahl aller möglichen Prognoseräume geteilt, hier: 15 / 500 = 0,03.

Die Berechnung des SAEI gestaltet sich demnach wie folgt:

$$(0{,}2-(0{,}03\cdot 0{,}15))/_{0{,}2} = {}^{0{,}1955}/_{0{,}2} = 0{,}9775$$

Der SAEI kann als Maßzahl interpretiert werden, die einen Mittelweg zwischen der maximal erreichbaren Genauigkeit (Trefferrate) und Effizienz (Abdeckung der Prognosegebiete) von Predictive Policing finden soll. Obwohl kein Wertebereich angegeben werden kann (dieser wird von der Achievable Efficiency bestimmt, die von Prognose zu Prognose unterschiedlich ist), geben höhere Werte an, dass die bewertete Prognose genauer und wahrscheinlich effizienter ist. Der exakte Zusammenhang bleibt jedoch unklar.

Die Konfusionsmatrix

Ein klassisches Qualitätsmaß zur Bewertung der Qualität automatischer Vorhersagemethoden, und in Teilen bei Umsetzungen von Predictive Policing anzutreffen, ist die Konfusionsmatrix oder auch Wahrheitsmatrix. Mit einer Vierfelder-Tabelle lassen sich übersichtlich Falsch Positive, Falsch Negative, Richtig Positive und Richtig Negative darstellen und vergleichen. Anhand des vorgenannten Beispiels würde dies, wenn in 100 von den 500 möglichen Prognosegebieten etwas passiert ist (je Gebiet ein WED), wie folgt aussehen:

		Delikt aufgetreten	
		Ja (Richtig)	**Nein (Falsch)**
Prognose	**Pos.**	15	5
	Neg.	85	395
	∑	100	400

Richtig Positive: Gebiete, die als Prognosegebiete vorausgesagt wurden und in denen etwas passiert ist (15 von 20).
Falsch Positive: Gebiete, die als Prognosegebiete vorausgesagt wurden und in denen nichts passiert ist (5 von 20).
Richtig Negative: Gebiete, die nicht als Prognosegebiete vorausgesagt wurden und in denen etwas passiert ist (85 von 480).
Falsch Negative: Gebiete, die nicht als Prognosegebiete vorausgesagt wurden und in denen nichts passiert ist (395 von 480).

Durch Bildung von entsprechenden Raten (z. B. Falsch-Positiv-Rate) kann sodann das Modell vergleichend analysiert werden. Auf die fast unzählig vorhandenen Möglichkeiten der Generierung von solchen Raten (und Indexwerten) wird in diesem Kontext aber verzichtet, da diese im Grundmodell konsistent immer die Prognose-

[7] Der RRI vergleicht die Anzahl der Delikte der aktuellen Vorhersageperiode mit der Vorgängerperiode.
[8] Public Engines ist ein ehemaliger, kommerzieller Anbieter für Datenanalysen und wurde im Jahr 2015 von Motorola Solutions übernommen.

dauer und den Prognoseraum einbeziehen, hiervon abhängen und so unabhängig von der Auslegung auch von der nachfolgenden Diskussion zur Variabilität und Validität betroffen sind.

Weitere Qualitätsmaße

Die Fallzahldifferenz

Neben den zuvor genannten Qualitätsmetriken, die teilweise recht komplex Prozente oder Indexwerte berechnen, ist es nicht unüblich, die Prognosegüte über die Fallzahldifferenz zu bestimmen. Durch einen Vergleich der Fallzahlen vor und nach Implementierung von Predictive-Policing-Umsetzungen wird durch einfache Subtraktion berechnet, ob durch den Einsatz dieser Methoden eine Fallzahlreduzierung – und damit ein Erfolg und Nutzen der eingesetzten Software – feststellbar ist. Die „Metrik" fokussiert folglich nicht einzelne Kriminalitätsprognosen, sondern versucht, den Effekt von Predictive Policing als Ganzes zu erfassen.

Dass diese Art und Weise der methodischen Überprüfung nicht vollständig valide ist, sei vorweggenommen. Korrelation ist nicht Kausalität und letztere lässt sich hier nie allein mit der Fallzahldifferenz feststellen. Wenn tatsächlich ein Fallzahlenrückgang nach dem Einsatz einer Predictive-Policing-Umsetzung zu verzeichnen ist, bleibt offen, ob dies auch kausal darauf zurückzuführen ist. Zum einen ist Kriminalität immer multikausal bedingt, zum anderen ist mit dem Einsatz von Predictive Policing immer „Verhalten" vonseiten der Polizei verbunden. Dies kann bewusst, beispielsweise ausgelöst durch die Erstellung einer Kriminalitätsprognose, aber auch unbewusst, ausgelöst durch externe, nicht beinflussbare Ereignisse, erfolgen. Letzteres ist sogar die Regel, da die Polizei als Institution grundsätzlich in Reaktion auf Notrufe oder Notlagen agiert. In diesem Kontext bleibt sodann offen, ob ein Fallzahlenrückgang auf bewusstes Verhalten der Polizei gegenüber erstellten Kriminalitätsprognosen zurückzuführen ist oder ob andere Ereignisse hierfür maßgeblich gewesen sind, auf welche die Polizei keinen Einfluss hat, z. B. Großschadensereignisse, politische Sonderlagen oder Ereignisse außerhalb des polizeilichen Kontextes. In der Folge sind Aussagen zur Wirksamkeit von Predictive Policing, ausschließlich reduziert auf einen Fallzahlenrückgang im Beobachtungszeitraum, mit besonderer Vorsicht zu betrachten. Sofern die Umsetzung begleitend durch andere Evaluationsdesigns, beispielsweise mit Methoden der qualitativen empirischen Sozialforschung, begleitet wird, kann möglicherweise konstatiert werden, dass es Anzeichen für einen reduzierenden Effekt des gewählten Ansatzes auf die Kriminalitätsentwicklung gibt. Da die nachfolgend zu diskutierende Variabilität von Qualitätsmetriken nicht auf die Fallzahlendifferenz zutrifft, wird auf diese im Weiteren auch nicht mehr eingegangen.

Experimentelle Designs

Es gibt einige Quellen, die sich der Bewertung der Qualität von Maßnahmen, basierend auf Methoden des Predictive Policing, mit klassischen experimentellen Methoden aus der Statistik annähern (vgl. Hunt 2014; Taylor et al. 2010). Dazu wird das betrachtete Gebiet in Kontroll- und Experimentalbereiche, die methodisch zwingend per Zufall ausgewählt werden müssen, aufgeteilt. Der Experimentalbereich ist Gegenstand von Maßnahmen basierend auf Predictive Policing, das Kontrollgebiet dagegen nicht. Zum Nachweis der Wirksamkeit wird nach einem festgelegten Zeitraum die Entwicklung bestimmter Indikatoren (z. B. Fallzahlen) im Experimentalgebiet den Indikatoren aus dem Kontrollgebiet gegenübergestellt und auf statistische Unterschiede geprüft. In klassisch kontrollierten Experimenten („Laborexperimente"), in denen die relevanten Außeneinflüsse kontrolliert und auch gezielt ausgeschaltet werden können, entspricht diese Methodik guter wissenschaftlicher Praxis (vgl. Eifler 2014: 202).

Hinsichtlich der Übertragbarkeit auf Predictive Policing kommen jedoch Zweifel auf. Die Zweifel gründen insbesondere auf der Annahme, dass die Indikatoren der betrachteten Gebiete bei Predictive-Policing-Umsetzungen in der Regel nicht vergleichbar sind. Streng genommen ist dies nur der Fall, wenn deren Struktur hinsichtlich Städtebau, Infrastruktur (wie Anschluss an das Autobahnnetz oder Fernverkehrsbahnhöfe), ebenso wie der sozioökonomischen Zusammensetzung der lokalen Bevölkerung (u. a. Bildungsgrad, Einkommen und Wohnverhältnisse) übereinstimmt. Der Nachweis dieser umfassenden Homogenität ist praktisch unmöglich zu führen, geschweige denn für Experimente in der Echtwelt herzustellen, denn sie müssten idealerweise methodisch per Zufall ausgewählt worden sein[9].

Neben diesem Problem führen auch die allgemeinen polizeilichen Maßnahmen und deren unmittelbare Auswirkungen zu einer Inhomogenität der betrachteten Gebiete, die es zur validen Be- und Auswertung des Experimentes unbe-

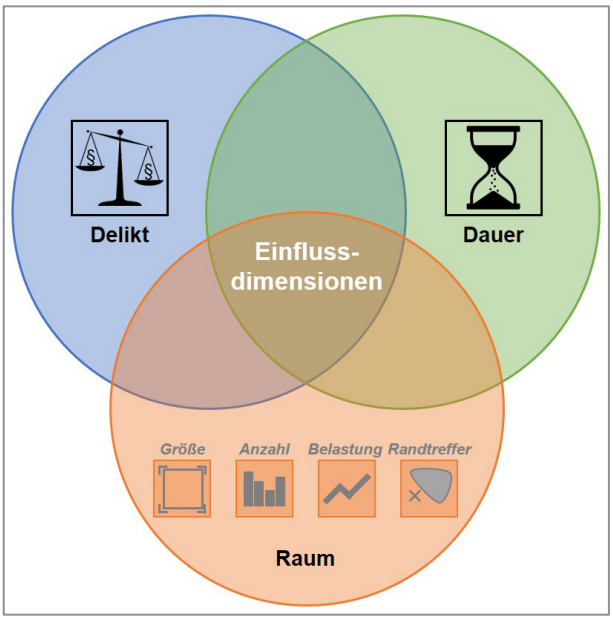
[9] Darüber hinaus ist dies aus forschungs-ethischen Gründen problematisch, da erstellte Kriminalitätsprognosen einer bestimmten Gruppe wissentlich vorenthalten werden.

dingt zu vermeiden gilt. Dreht man die Reihenfolge der Experimente um, also führt die Erfassung der Vergleichsindikatoren in dem betrachteten Gebiet durch, bevor Predictive Policing angewendet wird, besteht weiterhin eine Grundvariabilität, bedingt durch Interaktion der Polizei, der Einwohner sowie aller möglichen Arten von Umwelteinflüssen. Diese lässt sich nicht aufzeichnen, reproduzieren oder sinnvoll kontrollieren und schränkt damit die Vergleichbarkeit der betrachteten Zeiträume, unabhängig von deren struktureller Inhomogenität, schon vorab ein.

Variabilität und Validität der Qualitätsmetriken

Fast schon bahnbrechend lesen sich die Trefferraten, die von Polizeien oder Herstellern bei den softwaretechnischen Umsetzungen für das jeweilig genutzte oder beworbene Produkt proklamiert werden. So berichtet beispielsweise die Züricher Polizei von Trefferquoten über 80% (vgl. Golem 2014; Focus 2015). Mit dem Wissen der konkreten Berechnung der unterschiedlichen Qualitätsmetriken scheint die Einordnung solcher Ergebnisse zunächst leicht. Ein reflektiver Anstoß lässt jedoch erkennen, dass alle hier dokumentierten Qualitätsmetriken von drei wesentlichen Einflussdimensionen beeinflusst werden (Abbildung 4).

Abbildung 4: Einflussdimensionen



Mit dem Einbezug dieser Dimensionen in die Qualitätsmetriken wird nachfolgend dargestellt, dass sich entsprechende Ergebnisse sehr verschieden berechnen lassen und sich dadurch eine Variabilität in den Qualitätsmetriken bei Predictive Policing manifestiert. Diese Variabilität hat wiederum Auswirkungen auf die Validität der angewandten Metriken[10]:

Prognose-Delikt

Predictive Policing erfordert immer eine Zielvariable, die das Ziel der raum-zeitbezogenen Vorhersage darstellt. In der Regel bieten sich hier solche Kriminalitätsphänomene an, die nicht statisch sind (sonst wären es Hotspot-Analysen) oder der Kontrollkriminalität zuzurechnen sind (z. B. Betäubungsmittelkriminalität).

Grundsätzlich ist durch die Auswahl und Festlegung eines bestimmten Prognose-Delikts keine Variabilität bei sich anschließenden Qualitätsmaß-Metriken zu erwarten. Eine systematische Verzerrung entsteht jedoch, wenn Delikte, welche nicht prognostiziert wurden, als Treffer im vorausgesagten Prognosegebiet gewertet werden, wie beispielsweise ein Keller- oder Gewerbe-Einbruch bei einer Wohnungseinbruchprognose. Um eine Vergleichbarkeit von Prognoseergebnissen zu gewährleisten, müsste die Beantwortung solcher Fragen einheitlich erfolgen. Allerdings sind Bestrebungen einheitlicher Validierungsindizes derzeitig noch keine gängige Praxis im Bereich Predictive Policing. Die inhärente Subjektivität bei der definitorischen Festlegung auf eine bestimmte Auslegung lässt einen Vergleich zu anderen Umsetzungen von Predictive Policing nicht zu. Es darf folglich nur das Delikt in das Qualitätsmaß einbezogen werden, welches auch prognostiziert wurde. Andere Auslegungen verbieten sich im Hinblick auf ihre hohe Variabilität und die dadurch nicht zu erreichende Validität. Anderenfalls werden die Maßzahlen verzerrt und stellen fehlerhafte und in keinem Fall verwertbare Zahlen dar.

Prognose-Dauer

Eine Kriminalitätsprognose wird immer für einen bestimmten Zeitraum erstellt. Dieser Zeitraum kann, je nach Deliktsphänomen und polizeilichem Bedarf oder Nutzen, unterschiedlich sein. Denkbar, und entsprechend umgesetzt werden Tages- oder Wochenprognosen.

Mit Bezug zu den vorher dargestellten Qualitätsmetriken wirkt die Prognose-Dauer aber auch auf die Metrik und folgt dabei einem simplen Grundprinzip:

[10] Die Einflussdimensionen werden nachfolgend isoliert dargestellt. Sie wirken aber stets gemeinsam auf die zu berechnende Qualitätsmetrik.

Abbildung 5: Illustration der Auswirkung variabler Prognose-Dauern auf die Hit Rate
Prognosegebiete sind blau dargestellt, Gebiete mit Treffern rot, Wohnungseinbrüche gelb.

a) Prognose-Dauer: Ein Tag. Hit Rate: 0%

b) Prognose-Dauer: Eine Woche. Hit Rate: 33,33%

c) Prognose-Dauer: Zwei Wochen. Hit Rate: 50,00%

Je kürzer die Prognose-Dauer, je schlechter das Qualitätsmaß. Die darin enthaltende Logik ist denkbar einfach: Je länger der Zeitraum, auf den sich die Prognose bezieht (Gültigkeit), desto größer ist die Wahrscheinlichkeit, dass entsprechende Delikte in dem vorhergesagten Prognosegebiet auftreten können.

Abbildung 5 zeigt dies exemplarisch an einem Stadtteil für eine Tagesprognose, eine Wochenprognose und eine 2-Wochen-Prognose, welche alle am selben Tag beginnen und sich auf dieselben Prognosegebiete beziehen. Erwartungsgemäß ist der absolute Wert der Hit Rate für die Prognose-Dauer eines Tages (Abbildung 5 a) am schlechtesten. Die Wahrscheinlichkeit, dass ein Delikt in den Prognosegebieten auftritt, steigt mit der Anzahl der Delikte über die Zeit (Prognose-Dauer). So steigt sie in Abbildung 5 b) auf 33%, in Abbildung 5 c) auf 50%.

Werden folglich Trefferraten von Polizeien oder Herstellern angegeben, ist es essentiell wichtig zu wissen, für welche Prognose-Dauer die Kriminalitätsprognosen erstellt wurden. Eine Einordnung der Ergebnisse ist sonst nicht möglich und macht deren valide Bewertung schwierig.

Wie bei der Betrachtung zum Prognose-Delikt gilt: Qualitätsbewertungen müssen immer auf dieselbe Prognose-Dauer bezogen sein. Allenfalls variieren die zu Beginn eingeführten Prädispositionsfaktoren und führen zu nicht interpretierbaren Maßzahlen.

Prognose-Raum

Die Dimension des Prognose-Raums ist die im Kontext von Qualitätsmetriken, komplexeste Dimension, da sie sich in vier wesentliche Sub-Dimensionen aufteilt (Abbildung 4, untere Dimension: Raum):

Größe der Prognosegebiete

Je größer das Prognosegebiet, je höher die Hit Rate (oder der Indexwert). Denn mit wachsender Größe des Bezugsraumes steigt die Chance, dass im prognostizierten Gebiet das erwartete Ereignis eintritt. Im Umkehrschluss: Je kleiner die Größe des Bezugsraumes, desto unwahrscheinlicher wird es, dass das erwartete Ereignis im prognostizierten Gebiet auftritt.

Anzahl der Prognosegebiete

Neben der Sub-Dimension *Größe der Prognosegebiete* beeinflusst die Anzahl der Prognosegebiete auch die zuvor dargestellten Qualitätsmetriken in ihrer Variabilität und Validität: Je mehr Prognosegebiete herausgegeben werden, desto größer ist die Gesamtfläche der Prognosegebiete und folglich umso höher ist die Hit Rate bzw. der berechnete Indexwert. Auch hier ist es leicht nachvollziehbar, dass die Trefferwahrscheinlichkeit steigt, wenn mehr Prognosegebiete und da-mit potenzielle Trefferflächen bei einer Kriminalitätsprognose herausgegeben werden.

Mit der Herausgabe der Prognosegebiete an die operativen Polizeikräfte konstituiert sich darüber hinaus aber ein weiteres Validitätsproblem. Für die praktikable Umsetzung von Predictive Policing muss aus den berechneten Prognosegebieten eine Auswahl getroffen werden. Diese Auswahl kann anhand einer fixen Anzahl von Prognosegebieten oder aber anhand eines bestimm-

ten Schwellenwertes zur erwarteten Wahrscheinlichkeit (raum-zeitlicher-Dispositionsfaktor) erfolgen. Mit der Festlegung auf eine fixe Anzahl besteht das Risiko, dass bestimme Prognosegebiete nicht herausgegeben werden, die unter Umständen die gleiche oder eine annähernd gleiche Wahrscheinlichkeit aufweisen, aber durch die Fixierung aus der weiteren Betrachtung fallen. Eine inhaltliche Bewertung findet nicht statt. Die Entscheidung, welche Prognosegebiete Gegenstand des Predictive-Policings sein sollen, ist damit gewissermaßen willkürlich, da sie sich lediglich an der festgelegten Anzahl an Prognosegebieten orientiert.

Diese Problematik kann umgangen werden, indem die Festlegung anhand eines individuell zu bestimmenden Schwellenwertes erfolgt, mit der Folge, dass von Prognose zu Prognose eine unterschiedliche Anzahl an Gebieten herausgegeben wird. Mit der Herausgabe von einer unterschiedlichen Anzahl von Prognosegebieten ist aber eine zuverlässige, vergleichende Anwendung der hier genannten Qualitätsmetriken nur schwer möglich.

Belastung der Prognosegebiete

Die kriminelle Belastung bestimmter geografischer Räume einer Stadt (sog. Hotspots) hat ebenfalls großen Einfluss auf die Hit Rate oder den Indexwert. Denn werden Kriminalitätsprognosen auf hoch belastete Gegenden gelegt, steigt erwartungskonform auch die Wahrscheinlichkeit, dass es dort zu weiteren Delikten kommt. Der Aussagegehalt solcher Prognosen ist vergleichsweise gering und wird daher in der polizeilichen Praxis häufig kritisiert, da solche Erkenntnisse zuvor schon durch die Verwendung von Stecknadelkarten aus den siebziger Jahren gewonnen werden konnten. Eine Kriminalitätsprognose auf Hotspots steigert demnach zwar die Hit Rate oder den Indexwert, offenbart zugleich aber das darin enthaltene Dilemma: Die Kriminalitätsprognosen sind statisch und von wenig neuem Erkenntnisgewinn. Sie bilden das ab, was polizeilich bekannt ist. Aus methodischer Sicht handelt es sich um Hotspot-Analysen und nicht um Predictive Policing.

Randtreffer an Prognosegebieten

Die letzte Sub-Dimension ist die der sog. Rand- oder Beinahe-Treffer und deren Einbezug in die Bewertung im Berechnungsmodell. Abbildung 6 zeigt beispielhaft, wie Randtreffer auf einer geografischen Ansicht mit quartierbasierten Prognosegebieten auftreten können.

Abbildung 6: Illustration Randtreffer
Schraffiert ist das Prognosegebiet, gelbe Kreuze markieren Wohnungseinbrüche.

Die drei Wohnungseinbrüche liegen direkt neben dem Prognosegebiet. Aus prädiktiver Sicht also „knapp verfehlt“. Da Straftäter sich aber nicht an die künstlichen Grenzen eines Prognoseraumes halten (unabhängig davon, ob es Wohnquartiere oder Gitterzellen sind), sondern sich am natürlichen Raum und dessen Zusammensetzung orientieren, liegt es nahe, dass der ein oder andere Analyst oder Softwarehersteller solche Randtreffer ebenfalls noch in die eigene Trefferrate einbeziehen möchte (vgl. Public Engines 2014: 7).

Auf den ersten Blick ist das Problem der Randtreffer hauptsächlich eines der Variabilität. Unmittelbar daran anschließend stellt sich jedoch im Kontext einer Qualitätsbewertung von Predictive-Policing-Methoden die Frage der Auswirkung dieser Variabilität auf die Validität von unterschiedlichen Wertungsmethoden bei Randtreffern, beispielsweise:

- Sollen Randtreffer gezählt werden, sodass diese in Qualitätsmetriken wie dem PAI oder in der Betrachtung der Richtig Positiven bzw. Falsch Negativen Eingang finden?
- Welche Wertung der Randtreffer ist valide und damit zur sinnvollen und wahrheitsgetreuen Bewertung der Qualität der Umsetzung des Predictive Policing-Prozesses geeignet?
- Sollte ein Randtreffer als ein Viertel, ein halber oder als ein ganzer Treffer gewertet werden?

Je nachdem wie die Modellierung und die Prognoseberechnung erfolgt (Schritt 2 und 3 des Predictive-Policing-Prozesses), variieren folglich die Kriminalitätsprognosen und zeigen dadurch erneut eine stark vorhandene Variabilität in den Qualitätsmetriken. Die Dimension des Prognose-Raumes ist durch ihre Komplexität nachvollziehbarerweise am prominentesten in den hier vorgestellten Qualitätsmetriken.

Zusammenfassung und Ausblick

Abschließend bleibt anzumerken, dass durch jegliche Form von Variabilität in Qualitätsmetriken deren Vergleichbarkeit und Validität grundsätzlich in Frage gestellt werden muss. Ziel von Maßzahlen bei Predictive-Policing-Umsetzungen ist es, die Qualität der Modelle gesichert zu bewerten. Auf Grund der zuvor dokumentierten Variabilität ist eine solche sichere Bewertung allerdings nicht möglich. Eine Validität von Qualitätsmetriken ist für die hier dokumentierten und gängigen Metriken im Bereich von Predictive Policing dadurch nicht gegeben. Mit hoch variablen, subjektiven und uneinheitlichen Qualitätsmetriken lässt sich dies nicht realisieren.

Die Diskussion der Qualitätsmetriken, basierend auf den vorgestellten drei Einflussdimensionen bzw. deren Subdimensionen, dokumentiert, wie stark variierend die Metriken im Bereich von Predictive Policing sind und wie fragwürdig und wenig valide darauf basierende Aussagen zur Qualität der einzelnen Predictive-Policing-Umsetzungen sind. Dieses Problem beschränkt sich dabei nicht nur auf einfache Qualitätsmetriken wie die Hit Rate, sondern, bedingt durch die verschiedenen Einflussdimensionen, auch auf die komplexeren Qualitätsmetriken wie den PAI oder SAEI. Auch zukünftig noch zu entwickelnde Qualitätsmetriken müssen sich an diesen Einflussdimensionen messen.

Quantitativ basierte Bewertungen von Predictive-Policing-Umsetzungen sollten immer unter Vorbehalt interpretiert werden. Das betrachtete Phänomen ist zu komplex, um mit einer Maßzahl wie „80%“ oder „0,75“ adäquat beurteilt werden zu können. Besonders problematisch sind hierbei Vergleichsbetrachtungen, die schon in derselben Stadt für unterschiedliche Prognose-Zeiträume oder variable Prognose-Räume keine Allgemeingültigkeit und damit hohe Aussagekraft haben können.

Darüber hinaus konstituiert sich ein inhärentes Paradoxon: Die Anzahl der aufgetretenen Delikte sind Grundlage für die Trefferratenberechnung, obwohl die Delikte von der Polizei u. a. durch Predictive Policing aktiv verhindert werden sollen. Es wird also versucht etwas zu messen, was aktiv und in unbekanntem Maße von der Polizei beeinflusst wurde. Oder anders formuliert: Es soll etwas gemessen werden, was eigentlich verhindert werden soll und eventuell auch tatsächlich durch verstärkten Polizeieinsatz in den Prognosegebieten verhindert wird.

Felix Bode ist Mitarbeiter der Kriminalistisch-Kriminologischen Forschungsstelle im LKA NRW.
Kontakt: felix.bode@polizei.nrw.de
Web: http://www.lka.nrw.de

Florian Stoffel ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Datenanalyse und Visualisierung an der Universität Konstanz.
Kontakt: florian.stoffel@uni-konstanz.de
Web: http://infovis.uni.kn/~fstoffel

Prof. Daniel Keim ist Lehrstuhlinhaber des Lehrstuhls für Datenanalyse und Visualisierung an der Universität Konstanz.
Kontakt: daniel.keim@uni-konstanz.de
Web: http://infovis.uni.kn/~keim

Literaturverzeichnis

Balogh, Dominik (2016): Near Repeat-Prediction mit PRECOBS bei der Stadtpolizei Zürich. In: Kriminalistik, 5/2016, S. 335-341.

Braga, Anthony (2005): Hot spots policing and crime prevention: A systematic review of randomized controlled trials. In: Journal of Experimental Criminology, Vol. 1, S. 317-342.

Berk, Richard (2008): Forecasting Methods in Crime and Justice. In: Annual Review of Law and Social Science, H.1, Jg. 4, S. 219-238.

Berk, Richard; Sherman, Lawrence; Barnes, Geoffrey; Kurtz, Ellen & Ahlman, Lindsay (2009): Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. In: Journal of the Royal Statistical Society: Series A (Statistics in Society), H.1, Jg. 172, S. 191-211.

Box, George; Jenkins, Gwilym; Reinsel, Gregory & Ljung, Greta (2015): Time series analysis. Forecasting an Control, Fith Edition. New Jersey.

Chainey, Spencer; Tompson, Lisa & Uhlig, Sebastian (2008): The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. In: Security Journal, 21(1), S. 4-28.

Drawve, Grant (2014): A Metric Comparison of Predictive Hot Spot Techniques and RTM. In: Justice Quarterly, H. 3, Jg. 33, S. 369-397.

Eifler, Stefanie (2014): Experiment. In: Bauer, Nina & Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung, Wiesbaden, S. 195-209.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg & Xu, Xiaowei (1996): A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996, S. 226-231.

Fritsch, Dieter; Glemser, Michael; Klein, Ulrike; Sester, Monika & Strunz, Günter (1998): Zur Integration von Unsicherheit bei Vektor- und Rasterdaten. In: Geo-Informations-Systeme - Zeitschrift für raumbezogene Information und Entscheidungen, Vol. 11, Heft 4, S. 26-35.

Focus (2015): Vorhersage per Smartphone, Polizei-App sagt Ihnen wann der Dieb ins Haus kommt. Trefferquote bei über 80 Prozent. URL: http://www.focus.de/digital/handy/warte-auf-den-einbrecher-vorhersage-per-smartphone-polizei-app-sagt-euch-wann-der-dieb-ins-haus-kommt_id_4725470.html, zuletzt aufgerufen am 14.03.2017.

Golem (2014): Polizei nutzt Predictive-Policing Software gegen Einbrecher URL: https://www.golem.de/news/bayern-polizei-nutzt-predictive-policing-software-gegen-einbrecher-1408-108388.html, zuletzt aufgerufen am 14.03.2017.

Hart, Timothy & Zandbergen, Paul (2012): Effects of Data Quality on Predictive Hotspot Mapping - Final Technical Report. Washington. URL: https://www.ncjrs.gov/pdffiles1/nij/grants/239861.pdf, zuletzt aufgerufen am 14.03.2017.

Hunt, Priscilla; Saunders, Jessica & Hollywood, John (2014): Evaluation of the Shreveport Predictive Policing Experiment. Santa Monica. Rand Corporation.

Kass, Gordon (1980): An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 29, No. 2, S. 119-127.

Keim, Daniel; Kohlhammer, Jörn; Ellis, Geoffrey & Mansmann, Florian (2010): Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association. URL: http://www.vismaster.eu/book, zuletzt aufgerufen am 14.03.2017.

Kinkeldey, Christoph; MacEachren, Alan & Schiewe, Jochen (2014): How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. In: The Cartographic Journal, H. 4, Jg. 51, S. 372-386.

Levine, Ned (2008): The “Hottest” Part of a Hotspot: Comments on “The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime”. In: Security Journal, H. 4, Jg. 21, S. 295-302.

Mohler, George; Short, Martin; Malinowski, Sean; Johnson, Mark; Tita, George; Bertozzi, Andrea & Brantingham, Jeff (2015): Randomized Controlled Field Trials of Predictive Policing. In: Journal of the American Statistical Association, H. 512, Jg.110, S. 1399-1411.

Morgan, Millet; Henrion, Max & Small, Mitchell (1990): Uncertainity. Cambridge.

Motorola Solutions (Hrsg.) (2015): Predictive Analytics vs. Hotspotting. A Study of Crime Prevention, Accuracy and Efficiency. URL: https://www.motorolasolutions.com/content/dam/msi/docs/products/smart-public-safety-solutions/ilps/Predictive-Analytics-vs-Hotspotting-White-Paper.pdf, zuletzt aufgerufen am 31.01.2017.

Nexiga (2017): Geodaten auf höchster Ebene, URL: http://www.nexiga.com/geodaten-auf-hoechster-ebene, zuletzt aufgerufen am 14.03.2017.

Openshaw, Stan (1984): The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography, H. 38.

Perry, Walter; McInnis, Brian; Price, Carter; Smith, Susan & Hollywood, John (2013): Predictive Policing. The Role of Crime Forecasting in Law Enforcement Operations. Santa Monica. Rand Corporation.

Polizei Köln (2017): Wohnungseinbruchradar für Köln und Leverkusen, URL: https://www.polizei.nrw.de/koeln/artikel__13449.html, zuletzt aufgerufen am 14.03.2017.

Pollich, Daniela & Bode, Felix (2017 in Druck): Predictive Policing: Zur Notwendigkeit eines (sozial)wissenschaftlich basierten Vorgehens. In: Polizei & Wissenschaft.

Public Engines (2014): Predictive Analytics vs. Hot Spotting. A Study of Crime Prevention, Accuracy and Efficiency. URL: https://www.motorolasolutions.com/content/dam/msi/docs/products/smart-public-safety-solutions/ilps/Predictive-Analytics-vs-Hotspotting-White-Paper.pdf, zuletzt aufgerufen am 14.03.2017.

Saunders, Jessica; Hunt, Priscillia & Hollywood, John (2016): Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. In: Journal of Experimental Criminology, H. 3, Jg. 12, S. 347-371.

Silverman, Bernard (1986): Density Estimation for Statistics and Data Analysis. Chapman & Hall. London.

Taylor, Bruce; Koper, Christopher & Woods, Daniel (2010): A randomized controlled trial of different policing strategies at hot spots of violent crime. In: Journal of Experimental Criminology, H. 2, Jg. 7, S. 149-181.

Van Patten, Isaac; McKeldin-Coner, Jennifer & Cox, Deana (2009): A Microspatial Analysis of Robbery: Prospective Hot Spotting in a Small City. In: Crime Mapping: A Journal of Research and Practice, 1(1), S. 7-32.

Wang, Dawei; Ding, Wei; Lo, Henry; Stepinski, Tomasz; Salazar, Josue & Morabito, Melissa (2012): Crime hotspot mapping using the crime related factors – a spatial data mining approach. In Applied Intelligence, H. 4, Jg. 39, S. 772-781.

Zhang, Peter & Qi, Min (2005): Neural network forecasting for seasonal and trend time series. In: European Journal of Operational Research, H. 2, Jg. 160, S. 501-514.