CO-005

犯罪発生履歴データの機械学習による時空間カーネル密度推定型犯罪予測の最適化
Machine Learning Based Parameter Optimization of Spatiotemporal Kernel Density Estimation for Crime Forecasting

中川 淳子† 小西 勇介† 宮野 博義†
Junko Nakagawa Yusuke Konishi Hiroyoshi Miyano

1. はじめに

犯罪予測とは、犯罪の発生履歴データ等を用いて近い将来の犯罪発生場所を予測する技術で、警察のパトロール場所設定に活用され始めている。欧米で商用サービスが開始し、日本でも警察の運用事例がある [1]。

これまで犯罪予測の研究は、主に警察関係者や犯罪学の専門家により行われてきた[2][3]。犯罪の発生履歴は時間的・空間的に分布する離散データだが、実際の発生位置を含むデータは、専門家以外には入手困難だった。犯罪予測の従来手法の課題として、予測パフォーマンスを高めるためには、専門家が犯罪学の知見に基づいて、罪種や予測エリア等を考慮して統計量等を選択し、そのパラメータを設定する必要があるが、このような課題には、機械学習等の、データからパターンを見つけ出す手法によるアプローチが考えられる。2016 年、米国立司法省(National Institute of Justice：NIJ と略)は、犯罪予測手法の比較を目的としたコンテストを主催し、発生位置の緯度経度情報を含むデータを分析用に公開しており[4]、欧米を中心に研究開発が活性化している。

そこで本稿では、予測する犯罪発生場所を、発生履歴データから時空間カーネル密度推定で求める発生件数の密度分布の高密度エリアとし、パトロール可能な面積分の場所を予測した場合に、予測場所での発生件数が最大となるように、カーネル関数の最適バンド幅を機械学習で推定する手法を提案する。これにより、従来手法で必要だった専門家の犯罪学の知見に基づく統計量等の選択や、そのパラメータ設定が不要になる。NIJ の公開データを用いた実験で、提案手法で推定した罪種毎の最適バンド幅の傾向が、犯罪学の知見による傾向と合致することを示す。

以下、2 節で NIJ の犯罪予測コンテストを紹介し、3 節で従来手法の調査研究を紹介する。4 節で提案手法を説明し、5 節で NIJ の公開データを用いた実験について述べる。

2. NIJ 犯罪予測コンテスト

コンテストのタスクは「予測対象エリアをセルに分割し、所定の予測期間に犯罪発生が集中しそうな場所とそうでない場所を 1/0 で区別せよ」で、主な条件は以下である[4]。

・予測は以下の罪種毎に行う
　：侵入盗/路上犯罪/自動車盗/全パトロール通報
・予測期間は以下の 5 パターン
　：2017 年 3 月 1 日からの 1 週間/2 週間/1 か月間/2 か月間/3 か月間
・評価指標は以下から選択

†日本電気株式会社 NEC Corporation

Prediction Accuracy Index ：PAI=(n/N)/(a/A)
Prediction Efficiency Index ：PEI=PAI/PAI*

但し
n：予測できた発生件数　N：予測期間内全発生件数
a：予測したセル数　A：予測対象エリア全セル数
PAI*：n の代わりに発生件数が多いセル a 個での発生件数を用いた PAI

評価指標 PAI は予測セルの面積割合と予測できた件数割合の比、PEI は予測できた件数と、同じ予測セル面積割合で予測できる最大件数の比を意味する。公開データは米ポートランド警察の 5 年 3 か月分の 110 番通報の履歴で、各履歴は発生年月日、位置（緯度経度）、罪種で構成される。本データを地図上に表示したところ、交差点や路上への集中は見られないため、実際の発生位置に近いデータと考えられる。

なお、コンテスト応募資格者は米国内に限られ、提出期限は 2017 年 2 月末で結果はすでに発表されている。

3. 犯罪予測の従来手法

調査研究[2]では、網羅的な論文レビューに基づき、犯罪予測手法を、1)時空間クラスタの検出、2)犯罪の時空間的相互作用を考慮した犯罪発生の強度推定、3)環境要因からの犯罪発生リスクの予測、4)回帰分析による犯罪発生件数または確率の予測、の 4 つに分類している。さらに、調査研究[5]では、日本国内の車上狙いの犯罪発生履歴データを用いて、これら 4 つの各分類の代表的手法による、予測対象月の評価指標 PAI 等を評価している。

分類 1)は犯罪学の知見から適切な統計量を選択しそのパラメータを設定して、犯罪発生履歴データを用いて多発クラスタを計算する手法である。分類 2)は犯罪学の知見である犯罪の近接反復被害を取り入れたモデル式を作成し、パラメータである空間・時間バンド幅や近接反復被害の計算範囲を設定して予測を行う手法で、欧米警察の運用事例がある ProMap や PredPol の手法である。ここで、犯罪の近接反復被害とは、犯罪学で特に侵入盗について知られる知見で、「ある場所で犯罪が起こると、その『近隣』において、短い期間のうちに再び犯罪被害が起こること」であり、住宅侵入盗の発生後、少なくとも 2 週間の間は発生場所の近隣 200m において新たな住宅侵入盗発生のリスクが有意に高かったとしている[6][7][8]。

分類 3)、4)は犯罪発生に影響する要因を犯罪学の知見から選択しデータを収集して分析する手法で、要因とは人口密度や世帯の経済状況、道路の粗密や建築物の種類等である。3)の代表は米国警察の運用事例がある RTM(Risk Terrain Modeling)、4)は要因データと犯罪発生履歴データを用いて回帰分析を基に発生確率を予測する手法である。

これらの従来手法に共通する課題として、予測パフォーマンスを高めるためには、専門家が犯罪学の知見に基づいて、罪種や予測エリア等を考慮して統計量や要因を選択し、そのパラメータを設定する必要があることがあげられる。そこで、次節では、前記のような専門家による選択・設定を不要にする手法を提案する。

4. 提案手法

図 1 に提案手法の全体構成を示す。提案手法は、犯罪発生履歴データから時空間カーネル密度推定で求める発生件数の密度分布の高密度エリアを犯罪発生場所として予測する（4.1 節）。その際、予測に必要なパラメータであるカーネル関数のバンド幅を、犯罪発生履歴データに対する機械学習により推定する（4.2 節）。

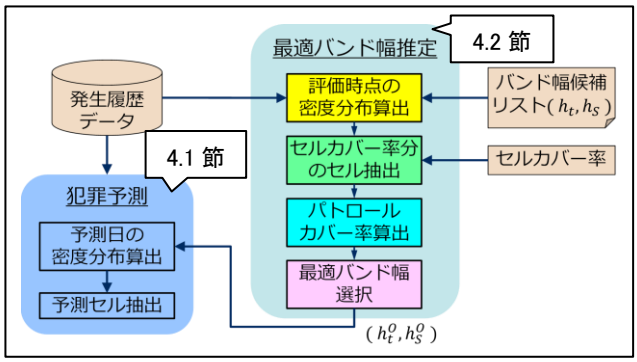

図 1　提案手法の構成

4.1 カーネル密度推定による犯罪予測

罪種毎の発生年月日と発生位置からなる犯罪発生履歴データ$(x_i,\ y_i,\ t_i), (i = 1,2,.,I)$を用いて、罪種毎に予測を行う。

予測対象エリアを同じサイズのセルに分割して、各セルの罪種毎の犯罪発生件数の密度分布を、時間成分と空間成分からなるカーネル関数で表す。年月日tにおける予測対象エリア内セル$(x,\ y)$での、ある罪種の犯罪発生件数の密度分布を（1）で定義する。

$$f(x,\ y,\ t) = \frac{1}{h_S^2 h_t} \sum_{i=1}^{I} K_t\left[\frac{t - t_i}{h_t}\right] K_S\left[\frac{x - x_i}{h_S},\ \frac{y - y_i}{h_S}\right] \quad (1)$$

（1）式は、3 節で説明した調査研究[5]の手法評価の比較対象用カーネル密度推定（但し空間成分のみ）を参考に作成した。K_t、K_Sは各々時間成分と空間成分のカーネル関数で、警察向け分析ツール CrimeStat[9]掲載のカーネル関数から、K_t は最も単純なトップハット（一様分布）を、K_S は[5]で用いた quartic（4 次関数）を選ぶ。セル中心と発生履歴との距離をx, yとすると以下である。

$$K_t(t) \propto 1 \qquad if \quad |t| < h_t$$
$$K_S(x,y) \propto \left(1 - \frac{x^2+y^2}{h_S^2}\right)^2 \qquad if \quad x^2 + y^2 < {h_S}^2 \quad (2)$$

図 2 にK_t、K_Sの分布を示す。このカーネル関数による（1）式の分布は、空間的にはバンド幅h_Sまで緩やかな減衰、時間的にはバンド幅h_tまで一定値である。

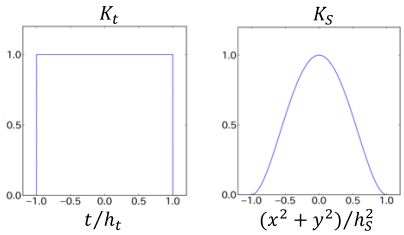

図 2　カーネル関数

年月日 t でのある罪種の発生の予測場所（セルの集合）は予測対象エリアから（1）式の値が高い高密度エリアのセルから順にをセルカバー率 β の割合だけ抽出したセルとする。セルカバー率とは、パトロールを実行するリソース（例えば人員数や車両数）でパトロール可能な面積の、予測対象エリア面積に対する割合で定義し、リソース量を考慮して設定する。

β ≡ (パトロール可能なセル数)
　／(予測対象エリア内全セル数)　(0 ≤β≤1)　（3）

4.2 機械学習による最適バンド幅推定

カーネル関数 K_t、K_Sのパラメータであるバンド幅（時間成分 h_t ・空間成分 h_S）は従来手法では専門家による設定が必要であった。これに対し提案手法は、発生履歴データ内に学習期間を設定して機械学習を行い、バンド幅候補の組の最適な組み合わせを推定する。図 3 はある学習期間における学習データ設定の説明図である。

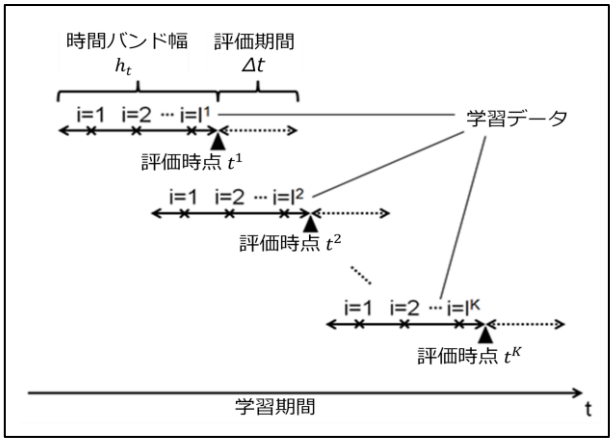

図 3　学習データ設定

学習期間内に評価時点t^k $(k = 1,2,..,K)$を設定し、t^kでの発生とみなす期間を評価期間Δtとする。

与えたバンド幅候補の組$(h_t,\ h_S)$毎に、ある学習期間の評価時点t^k $(k = 1,2,..,K)$での発生件数密度分布$f(x,\ y,\ t^k)$を時間バンド幅 h_t内の発生履歴データI^k件を用いて（1）式で算出し、$f(x,\ y,\ t^k)$の値が予測対象エリアの全セルの上位からセルカバー率 β 以内となるセルの集合 $G^{k(\beta)}$を抽出する。次にt^kの評価期間内で実際に犯罪が発生したセルと、t^kでの$f(x,\ y,\ t^k)$の高密度エリアのセルすなわち予測場所がどの程度一致するかを、パトロールカバー率を用いて評価する。

パトロールカバー率とは、セルカバー率（３）式分の面積を予測した場合の、予測対象エリア内全発生件数に対する、予測場所での発生件数割合、すなわち予測の的中率である。

パトロールカバー率

$$\equiv \sum_{k=1}^{K}(\,t^k \text{ の評価期間内発生のうち} G^{k(\beta)} \text{での発生数}) \Big/ \sum_{k=1}^{K}(\,t^k \text{ の評価期間内発生数}) \quad (4)$$

ある学習期間で全バンド幅候補組のうちパトロールカバー率を最大にする組を最適バンド幅(h_t^0, h_s^0)とする。

パトロールカバー率のセルカバー率に対する変化傾向は、一般的にバンド幅組により異なると考えられる（図 4）。

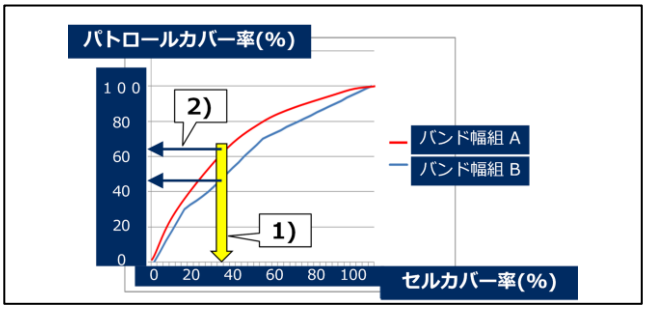

図 4　セルカバー率とパトロールカバー率

提案手法は 1)パトロールのリソース量に基づく予測面積割合（セルカバー率）を設定して、2)予測できた件数割合（パトロールカバー率）を最大にするバンド幅組を選ぶ。パトロールカバー率は、パトロール警察官がセルカバー率分のセル全域に評価期間中常駐する場合に犯罪発生に遭遇する件数割合を意味するので、パトロール時に遭遇する件数割合ではないが、パトロールリソースに対するパトロール効果の目安を最大にする点で、提案手法はパトロールの運用に役立つと考えられる。

5. 実験

2 節で説明した NIJ 犯罪予測コンテストの公開データを用いて、提案手法にバンド幅候補の値を与えて、セルカバー率 1%の場合の罪種毎の最適バンド幅を推定する。予測対象エリアであるポートランド署所管エリアは約 20km 四方で、セル幅は 75m とし、空間バンド幅はセル幅の 75m から 2000m まで、時間バンド幅は 7 日から 4 年までを候補値として与え、全組み合わせ数は 136 組である。学習期間は季節変動がある場合にその特徴を捉えるために 1 ケース 3 か月間とし、公開データ内で時間バンド幅 4 年を設定できる範囲に過去から順に 1~5 の 5 ケースを設定する。罪種は自動車盗、路上犯罪、侵入盗とする。

5.1 自動車盗

図 5 に、全学習期間における、バンド幅候補組毎のパトロールカバー率の相対値を、色で示す。学習期間毎の最適バンド幅は、時間バンド幅 2-4 年、空間バンド幅 75-100m の範囲だった。これは過去 2-4 年の発生場所から 75-100m という近距離で多く発生する傾向があることを意味する。犯罪学では自動車盗は駐車場での発生が多く、かつ多く発生する駐車場には照明が暗い等の条件があるとの知見がある[10]。駐車場の場所は数年間変わらないと考えると、実験結果は既存の知見と合致すると言える。

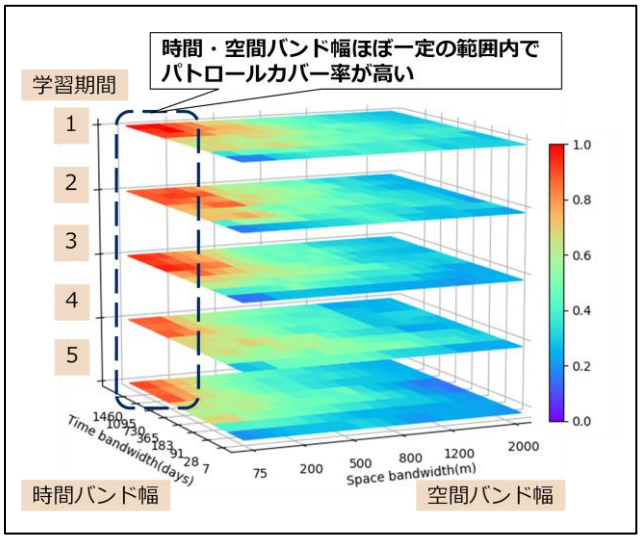

図 5　自動車盗のパトロールカバー率

5.2 路上犯罪

図 6 の学習期間毎の最適バンド幅は、時間バンド幅 1-2 年、空間バンド幅 75m で、全学習期間でほぼ変わらず、過去の発生場所の近くで発生する傾向が安定していた。また、時間バンド幅は 3 か月以上であればパトロールカバー率はほぼ変わらず（差は 5%以内）、発生傾向は時間バンド幅によらなかった。路上犯罪には複数の罪種が含まれ、暴行、強盗等の対人犯罪と騒乱等の秩序違反があるが、犯罪学では、暴行と飲酒店との関係、ひったくり等の路上強盗は都市部では公共交通機関の駅のそばで多い、若者の秩序違反の種類と多発場所の対応付け等の知見がある[10]。これより路上犯罪は特定の性質を持つ路上の狭いエリアで多発すると考えると、実験結果は既存の知見と合致すると言える。

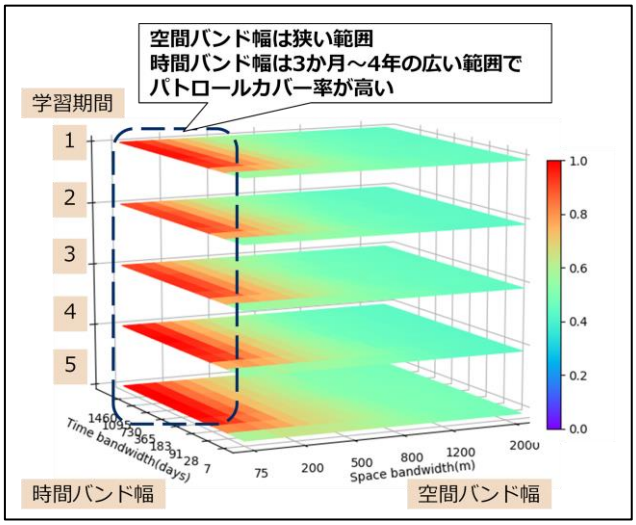

図 6　路上犯罪のパトロールカバー率

5.3 侵入盗

図 7 の最適バンド幅は学習期間で変化し、学習期間毎の最適バンド幅は、時間バンド幅は 3 年以上と半年、空間バンド幅は 150m と 500m だった。パトロールカバー率が相対的に高い時間バンド幅はおおむね半年以上だが、学習期間 4 では時間バンド幅 28 日～1 年、空間バンド幅 150m の範囲でパトロールカバー率が高く、前後の学習期間では同じ範囲でのパトロールカバー率は高くなかった。また、学習期間 2、4 ではパトロールカバー率が高い組み合わせが空間バンド幅 1km 以上まで広がっていた。

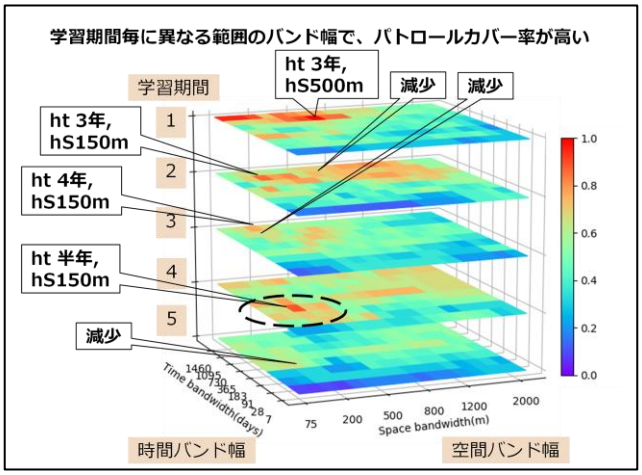

図 7　侵入盗のパトロールカバー率

侵入盗についての犯罪学の知見として、3 節で説明した犯罪の近接反復被害がある。学習期間 4 で、パトロールカバー率が高い時間バンド幅の範囲 28 日～1 年と空間バンド幅の値 150m は、近接反復被害による「発生後少なくとも 2 週間の間はその近隣 200m において新たな住宅侵入盗発生のリスクが有意に高かった」という知見の値と近い。よって既存の知見と合致する実験結果が得られたと言える。

一方、提案手法の犯罪発生件数密度分布の定義式（1）と今回選んだカーネル関数（2）は、空間的に近隣の発生を考慮し、時間的に時間バンド幅内だけの発生を考慮することを意味する（4.1 節）ので、近接反復被害の性質をモデル化したとも言える。そのため、侵入盗の発生が近接反復被害だけで説明できれば、学習期間に対しほぼ変わらないバンド幅が得られると考えられるが、実験結果はそうではなかった。提案手法のモデルは改善の余地がある。

（補足）表 1 は参考用の、提案手法でセルカバー率 1%の場合に推定した、罪種毎の最適バンド幅とパトロールカバー率、比較手法のパトロールカバー率である。比較手法は、調査研究[5]の手法評価の比較対象用カーネル密度推定で専門家が設定したバンド幅（6 か月, 250m）を、提案手法に与えた実験結果とした。両手法とも、最高のパトロールカバー率となる学習期間での値を記載した。調査研究[5]の手法評価では、車上狙いと部品狙いを合算した罪種につき、セル幅 25m で算出していて、提案手法の実験とは異なるが、提案手法による最適バンド幅推定値は調査研究[5]の設定値とは異なり、その結果パトロールカバー率は向上している。これらの最適バンド幅の将来の期間における有効性は今後検証したい。

罪種	提案手法(最適バンド幅)	比較手法 (6か月,250m)
自動車盗	14.8% (3年,100m)	9.38%
路上犯罪	34.2% (2年,75m)	26.2%
侵入盗	10.7% (3年,500m)	6.20%

表 1　罪種毎のパトロールカバー率

6. まとめ

本稿では、犯罪発生履歴データのみを用い、犯罪に関する知見を特に必要とせず専門家による設定が不要な犯罪予測手法として、予測する犯罪発生場所を、犯罪発生履歴データから時空間カーネル密度推定で求める発生件数の密度分布の高密度エリアとし、パトロール可能な面積割合の場所を予測した場合に、予測場所での発生件数が最大になるように、カーネル関数の最適バンド幅を機械学習で推定する手法を提案した。

NIJ 公開データを用いた実験で、提案手法で予測面積割合 1%の場合に推定した自動車盗、路上犯罪、侵入盗の最適バンド幅と、学習期間に対する安定性を検討した。その結果、自動車盗は過去数年間の発生場所から近距離での多発を示し、特定の条件の駐車場で多く発生するという犯罪学の知見と合致した。路上犯罪は 3 か月以上前の発生場所のごく近くだけでの多発を示し、ある種類の店舗や施設近辺で多く発生するという知見と合致した。侵入盗の最適バンド幅は学習期間で変化したが、犯罪の近接反復被害の知見によるバンド幅に近い値が得られた学習期間があり、この点で既存の知見と合致した。一方、提案手法の密度分布の定義式と選択したカーネル関数は近接反復被害の性質をモデル化したとも言えるが、学習期間に対し安定な最適バンド幅は得られなかったため、侵入盗の発生の説明には不十分なモデルと考えられ、今後改善を検討する。

参考文献

[1] 国立国会図書館 調査及び立法考察局,” 人工知能・ロボットと労働・雇用をめぐる視点：科学技術に関する調査プロジェクト報告書 第 2 部 分野別の動向, p.75, p.77 (2018). http://dl.ndl.go.jp/view/download/digidepo_11065186_po_20180405.pdf?contentNo=1
[2] 大山智也, 雨宮護,島田貴仁,中谷友樹, “地理的犯罪予測研究の潮流”, GIS-理論と応用, Vol.25, No.1 (2017).
[3] Walter L. Perry, Brian Mclnnis, Carter C. Price, Susan C. Smith, John S. Hollywood, Predictive Policing, RAND (2013)
[4] National Institute of Justice , “Real-Time Crime Forecasting Challenge” (2016). https://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx
[5] 大山智也, “日本における地理的犯罪予測手法の開発”, 公益財団法人日工組社会安全研究財団 2016 年度研究助成事業報告書 http://www.syaanken.or.jp/wp-content/uploads/2017/12/RP2016B_003.pdf
[6] 菊池城治, 雨宮護, 島田貴仁, 齋藤知範, 原田豊, “近接反復被害の罪種間比較-時空間 K 関数の応用-“, GIS-理論と応用, Vol. 18, No.2 (2010).
[7] リチャード・ウォートレイ, ロレイン・メイズロール, “環境犯罪学と犯罪分析”, (財)社会安全研究財団 (2010)
[8] Shane Johnson, Kate Bowers, Ken Pease, “Predicting Burglary Hotspots”, 2010 Conference on Evidence-Based Policing, University of Cambridge, Institute of Criminology http://www.crim.cam.ac.uk/events/conferences/ebp/2010/pbh.ppt
[9] CrimeStat http://www.nij.gov/topics/technology/maps/pages/crimestat.aspx
[10] Center for Problem-Oriented Policing, “Problem-Specific Guides” http://www.popcenter.org/problems/