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

中川 淳子　　西村 祥治　　宮野博義

富山県立大学　情報基盤工学講座
t815008@st.pu-toyama.ac.jp

April 23, 2021

本研究の背景

背景

犯罪の発生年月日と場所からなる犯罪履歴データ等を用いて, 将来どこで犯罪が発生するかを予測する犯罪予測という技術が, 警察の行うパトロール業務で利用されてはじめている.
米国立司法省研究所 (National Institute of Justice : NIJ) は,2016 年に犯罪予測手法をテーマとしたコンテストを主催した際に, 米ポートランド警察における犯罪発生位置の緯度経度情報を含む履歴データを分析用に公開したため, 従来と異なる分野の研究者による犯罪予測の研究活性化が期待されている.

TOYAMA Prefectural University

研究の目的

目的

本論文では, パトロール業務支援を目的に, 犯罪履歴データから, 近い将来の犯罪発生リスクの高い場所を予測する. まず既存の研究を紹介し, 次に本論文の提案手法の説明, 最後に提案手法を NIJ の公開している犯罪履歴データに適用し, 結果を犯罪学における既存の知見を参考にして考察する.

警察, 犯罪学専門家による既存研究

犯罪予測の分類

- 犯罪予測：犯罪の増加する場所、時間を予測
- 犯罪者予測：犯罪を行うリスクのある個人, グループを識別
- 犯罪被疑者の本人属性予測：犯罪被疑者のプロファイルを作成
- 犯罪被害者予測：犯罪被害者リスクのある個人, グループを識別

犯罪予測手法の分類

- 時空間クラスタの検出
- 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
- 環境要因からの犯罪発生リスクの予測
- 回帰分析による犯罪発生件数または確率の予測

異分野の研究者による既存研究

既存研究の分類

- 犯罪発生時間中心
- 犯罪発生場所中心　←本論文で提案される手法
- 犯罪者個人やコミュニティ中心

犯罪発生場所の分析手法:ホットスポット分析

- 分析対象エリアのなかで, 過去の犯罪発生が集中するエリアをホットスポットと呼ぶ. これを予測に活かす.
- ホットスポットの分析の手法としてグリッドマッピング, カーネル密度推定が挙げられる. 生成するグリッドセルの大きさ, カーネル関数のバンド幅 (パラメータ) は経験, 知見によって定められている.

既存研究の評価指標

ホットスポット分析の評価指標

- PAI(Prediction Accuracy Index):ホットスポット分析で求めたエリアが, どの程度犯罪場所を予測できているかを, 分析手法間で比較するための評価指標.
- 予測対象エリアに対する予測セルの面積割合と, 発生件数に対する予測できた件数割合の比

既存研究の課題

既存研究の課題

- 犯罪予測研究において, データ選択やパラメータ設定は経験・知見に基づいていて, 従来の専門家以外には設定が難しい.
- パトロール実施に割り当てられるリソース量を, 何らかの方法で制約したうえで PAI により予測精度を評価することが望ましい.

提案手法 (最適化 KDE 法)

概要

経験, 知見によるデータ・パラメータ設定を極力無くすため, カーネル密度推定により犯罪発生場所を予測し, そのパラメータである最適なバンド幅を犯罪発生履歴データの機械学習により推定する手法を提案する. そして, バンド幅推定の評価関数にパトロールリソース量の制約条件を導入する.

- 既存研究の手法と予測結果を合わせるために, 予測対象エリアを均一の形状のセルに分割し, セル単位で予測場所を出力する.
- パトロール実施可能なセルの, 予測対象エリアに対する面積割合をセルカバー率, その範囲内で選択した予測セルにおける予測的中率をパトロールカバー率と定義する.
- セルカバー率の制約下で, パトロールカバー率をカーネル関数のバンド幅に関して最大化する.

提案手法のブロック図

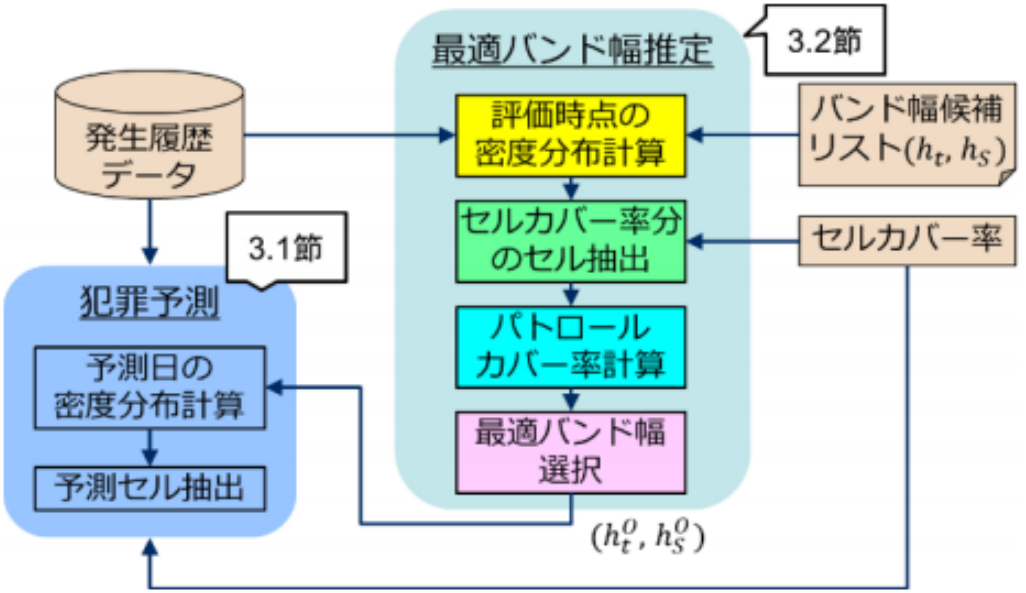


図 1: 最適化 KDE 法のブロック図

TOYAMA Prefectural University

時空間カーネル密度推定

犯罪発生位置と発生年月日からなる罪種ごとの犯罪発生履歴データ (x_i, y_i, t_i),$(i = 1, 2, ..., I)$ を用いて, 予測対象エリア内のセルの中心点 (x, y), 年月日 t における犯罪発生件数の密度推定値を時間成分と空間成分のカーネル関数を各々K_t, K_s として次式で定義する. また,h_t, h_s はそれぞれカーネル関数 K_t, K_s のパラメータのバンド幅である.

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

セルの中心点 (x, y) から犯罪発生位置 (x_i, y_i) への相対座標を (x_r, y_r), すなわち,$x_r = x_i - x$,$y_r = y_i - y$ とすると次式で表すことができる.

$$K_t(t) \propto 1 \quad \text{if } |t| < h_t$$

$$K_S(x_r, y_r) \propto \left(1 - \frac{{x_r}^2 + {y_r}^2}{h_S^2} \right)^2 \quad \text{if } {x_r}^2 + {y_r}^2 < {h_S}^2 \quad (2)$$

時空間カーネル密度推定

K_t, K_s の分布は図 1 に示される.

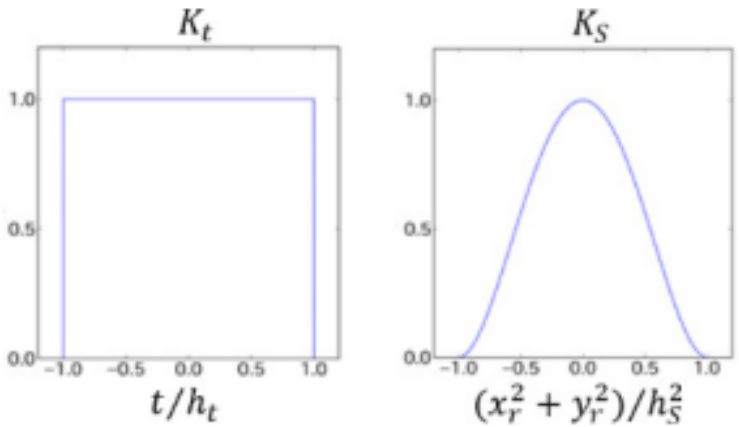


図 2: カーネル関数

また, セルカバー率は式 (3) で定義する面積割合である,

$$\begin{aligned} &\text{セルカバー率 } \beta \\ &\equiv (\text{パトロール可能なセル数}) \\ &\quad /(\text{予測対象エリア内全セル数}) \quad (0 \leq \beta \leq 1) \end{aligned} \tag{3}$$

機械学習による最適バンド推定

カーネル関数のパラメータであるバンド幅 (h_t, h_s) の候補リストを与え, 発生履歴データに含まれる年月日内に学習期間を設定して機械学習を行い, 候補リストの中から最適な組み合わせを選ぶ.
学習データ収集年月日として複数の評価時点 $t^k (k = 1, 2, ..., K)$ を設定し, バンド幅の候補リストから組を用いて評価時点 t^k における密度推定値 $f(x, y, t^k)$ を t^k から過去時間バンド幅 h_t 内の I^k 件の発生履歴データを学習データとして, 式 (1) で計算する.

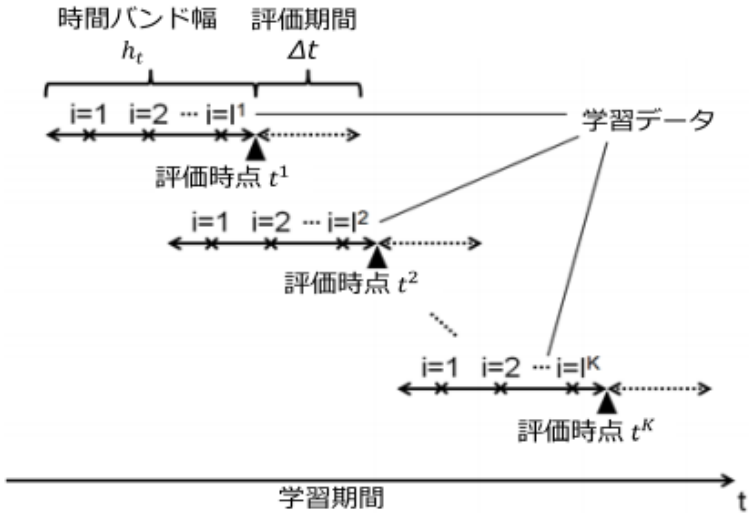


図 3: 学習データの設定方法

機械学習による最適バンド推定

$G^{k(\beta)}$:予測対象エリア内の全セルから,$f(x, y, t^k)$ が高い順に, セルカバー率 β 以内の割合のセルを抽出した t^k における発生を予測するセル集合　とすると, パトロールカバー率は次式によって表される.

$$
\begin{aligned}
&\text{パトロールカバー率} \\
&\equiv \sum_{k=1}^{K} (t^k \text{ の評価期間内発生のうち } G^{k(\beta)} \text{ における} \\
&\qquad \text{発生数}) \\
&/ \sum_{k=1}^{K} (t^k \text{ の評価期間内発生数}) \qquad (4)
\end{aligned}
$$

実験 (1)

実験概要

最適化 KDE 法を NIJ 犯罪予測コンテスト公開データに適用し, 自動車盗, 路上犯罪, 侵入盗の罪種ごとに予測を行う, 次に, グリッドマッピングを改良して同じ分析を行い, 結果を比較・考察をする.

コンテストのタスク

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.
予測は罪種ごとで, 予測期間は特定の年月日から 1 週間/2 週間/3 週間 など.

実験 (2)

条件設定

- 公開データ：米ポートランド警察の 5 年 3 か月分の犯罪発生履歴. 罪種, 発生年月, 発生場所の緯度経度.
- 予測対象エリア：発生が密集する矩形 (東西約 23km, 南北約 17km). セルの形状は 1 辺の長さ 75m の正方形.
- セルカバー率を 1%と 5%と設定する.

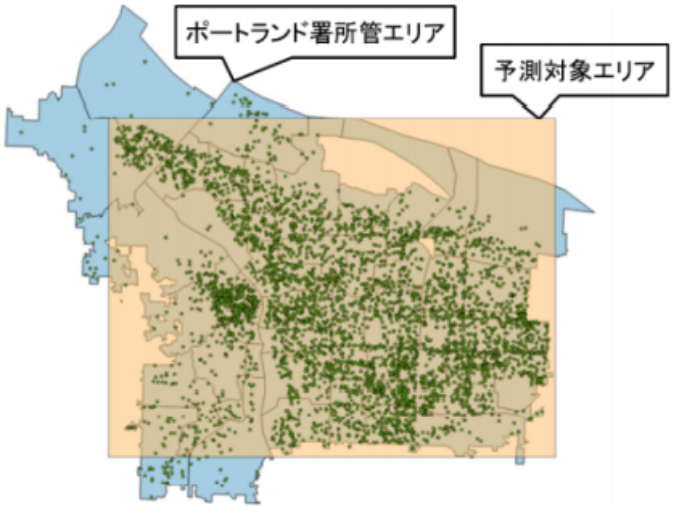


図 4: ポートランド署所管エリアと侵入盗発生地点

実験 (3)

罪種	予測対象エリア内件数	セルあたり件数
自動車盗	9,658 件	0.141 件
路上犯罪	163,785 件	2.395 件
侵入盗	5,277 件	0.077 件

（注）予測対象エリア内セル数　68,400.

Table 1: 罪種別犯罪データ数

種類	値	組数
時間 バンド幅(日)	7, 28, 91, 183, 365, 730, 1095, 1460	8 組
空間 バンド幅(m)	75, 100, 150, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000	17 組
組合せ合計		136 組

Table 2: バンド幅候補リスト

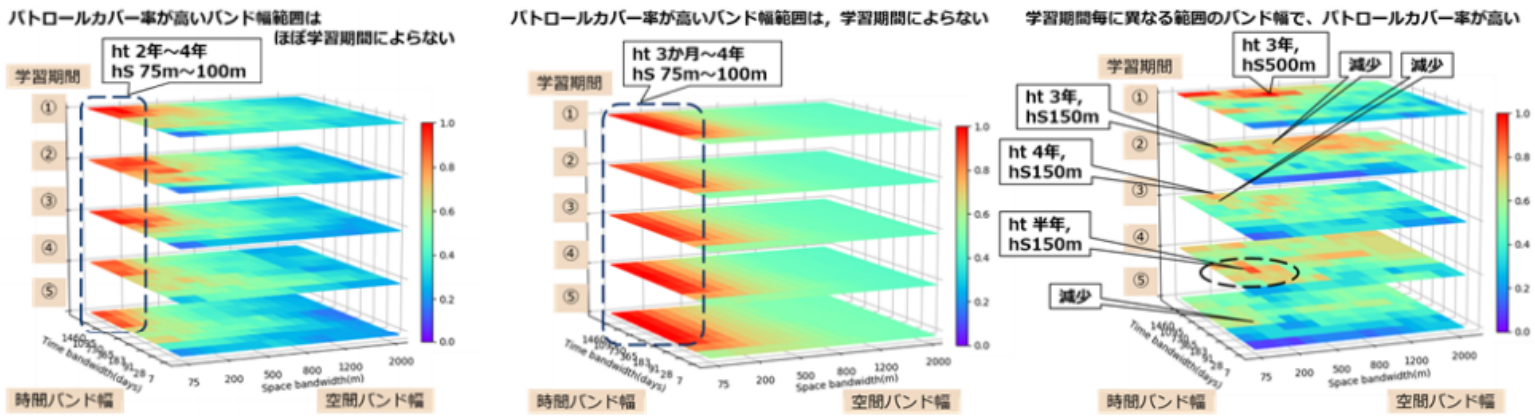
No.	学習期間	自動車盗	路上犯罪	侵入盗
①	2016 年 3 月-5 月	508 件	8,278 件	196 件
②	2016 年 6 月-8 月	506 件	9,709 件	228 件
③	2016 年 9 月-11 月	710 件	8,157 件	258 件
④	2016 年 12 月 -2017 年 1 月	821 件	6,739 件	251 件
⑤	2017 年 3 月-5 月	792 件	8,358 件	261 件

Table 3: 学習期間と，予測対象エリア内の発生データ数

実験 (4)

罪種ごとの実験結果（セルカバー率 1%）

- 自動車盗:過去 2〜4 年の発生場所から 75〜100m の距離で多く発生する.
- 路上犯罪:過去 3 か月〜4 年で 75〜100m の距離で発生している.
- 侵入盗:学習期間に対して変化し不安定である.



(a) 自動車盗　(b) 路上犯罪　(c) 侵入盗

図 5: バンド幅候補に対する学習期間ごとのパトロールカバー率

実験 (5)

罪種ごとの最適バンド幅組

セルカバー率 1%と 5%における最適空間バンド幅は, 自動車盗, 路上犯罪で狭く, 侵入盗では広い傾向がある.

罪種	学習期間	セルカバー率	
		1%	5%
自動車盗	①	(1095, 100)	(1460, 75)
	②	(1095, 100)	(1460, 150)
	③	(730, 75)	(1460, 75)
	④	(730, 75)	(1095, 75)
	⑤	(1460, 75)	(1460, 75)
路上犯罪	①	(365, 75)	(730, 75)
	②	(183, 75)	(730, 75)
	③	(365, 75)	(730, 75)
	④	(730, 75)	(730, 75)
	⑤	(730, 75)	(365, 75)
侵入盗	①	(1095, 500)	(1460, 300)
	②	(1095, 150)	(1460, 150)
	③	(1460, 150)	(1460, 300)
	④	(183, 150)	(1460, 400)
	⑤	(183, 150)	(1460, 300)

Table 4: 最適バンド幅組 (時間バンド幅 (日), 空間バンド幅 (m))

実験 (6)

改良型グリッドマッピング

任意のセルカバー率 β でパトロールカバー率を計算できるように, 同一順位セルがあった場合に, それらのなかからランダムに予測セルを抽出して, 合計 β 個のセルを予測セル集合 $G^{k(\beta)}$ とする.

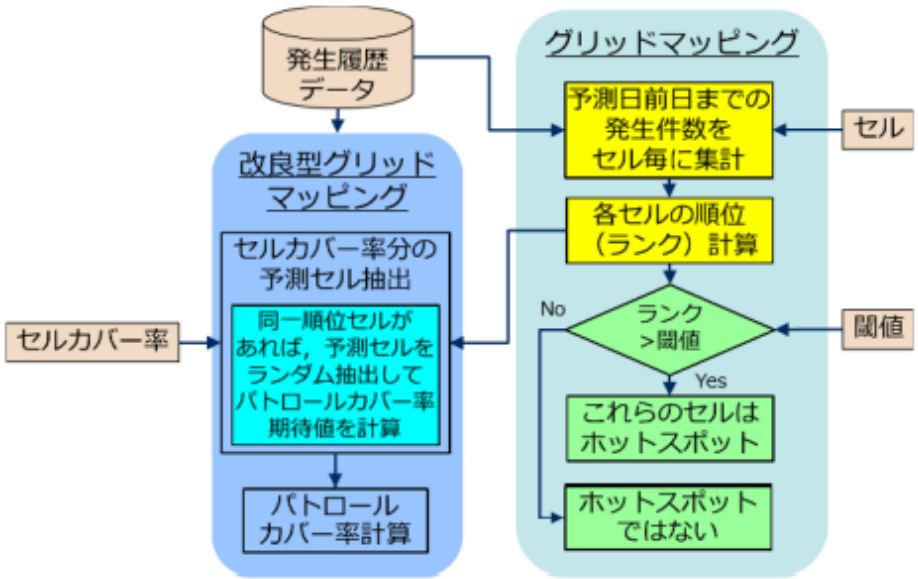


図 6: グリッドマッピングのブロック図

実験 (7)

最適化 KDE 法と改良型グリッドマッピングの比較

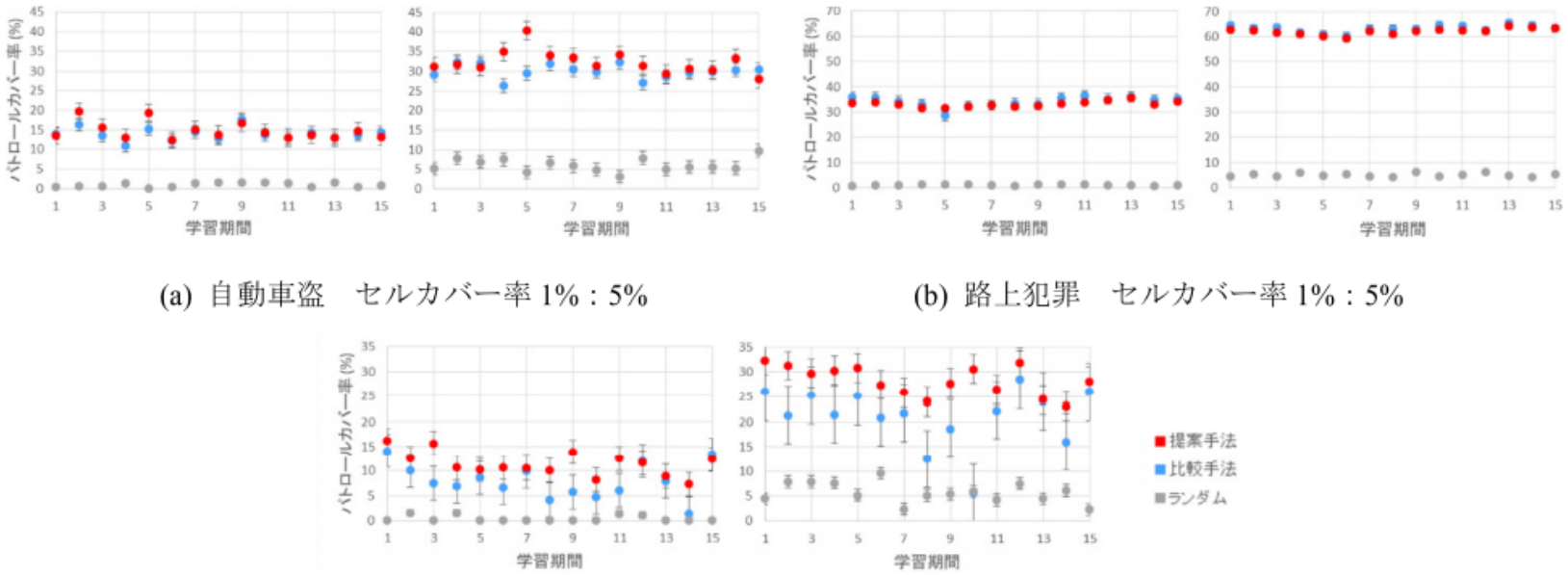
罪種	予測手法	セルカバー率		学習期間ごとの発生件数
		1%	5%	
自動車盗	提案手法	**14.72** ± 2.16	**32.35*** ± 2.82	222.47 ± 48.26
	比較手法	13.95 ± 1.58	29.97 ± 1.72	
	ランダム	0.95 ± 0.53	6.07 ± 1.62	
路上犯罪	提案手法	33.22 ± 1.13	62.04 ± 1.26	2749.40 ± 360.11
	比較手法	**34.17*** ± 1.98	**63.18**** ± 1.38	
	ランダム	1.07 ± 0.19	5.08 ± 0.64	
侵入盗	提案手法	**11.43**** ± 2.42	**28.26**** ± 2.89	79.60 ± 10.20
	比較手法	7.94 ± 3.40	21.05 ± 5.74	
	ランダム	0.37 ± 0.62	5.69 ± 1.20	

（注）平均値 ± 標準偏差，太字は罪種ごとの最高平均パトロールカバー率を示す．

Wilcoxon の符号付順位検定（片側） *：$p<0.0125$，**：$p<0.0005$

Table 5: 全学習期間の平均パトロールカバー率

実験 (8)



(a) 自動車盗　セルカバー率 1% : 5%

(b) 路上犯罪　セルカバー率 1% : 5%

(c) 侵入盗　セルカバー率 1% : 5%

図 7: 学習期間ごとのパトロールカバー率

実験 (9)

罪種	セルカバー率	提案手法	比較手法
自動車盗	1%	**14.72**	13.95
	5%	6.47	5.99
路上犯罪	1%	33.22	**34.17**
	5%	12.41	12.64
侵入盗	1%	**11.43**	7.94
	5%	5.65	4.21

(注) 太字は罪種ごとの最高 PAI を示す.

Table 6: 全学習期間の平均 PAI

まとめと課題

実験結果のまとめ

- 平均パトロールカバー率では, 比較手法と提案手法で 1 パターンを除いて優位差が認められ、6 つのうち 3 パターンが比較手法よりも高い.
- 学習時間とパトロールカバー率の散布図で, 提案手法は比較手法よりも高いパトロールカバー率を維持し, 値の変動幅も比較手法に比べ小さい.

課題

- 複数の時空間カーネル変数を用いた犯罪予測モデルへの拡張.
- 地理的, 時間的な環境要因データを用いる予測との統合的手法の開発,