明治大学総合数理学部

2021 年度

卒　業　研　究

インシデント損害金額を推定するウェブサイトの開発と評価

学位請求者　先端メディアサイエンス学科

伊藤　充司

目次

第 1 章

はじめに

情報化社会が加速していく中，不正アクセスによる情報漏洩などのサイバーインシデントが増加傾向にある．2021 年 4 月 28 日には，株式会社ネットマーケティングは提供する恋活・婚活マッチングアプリ「Omiai」の管理サーバに不正アクセスがあり、約 171 万件の会員情報が流出した [4].

事件を起こさないためには企業が情報漏洩に対するリスクを正確に認識し，適切な対策をする必要がある．本研究では，各企業のインシデントリスクをに定量化するために，専門知識のない一般の人でも簡単にインシデント被害金額を推定できるウェブサイトの開発をした．本サイトを用いて 2021 年現在のインシデントデータで評価した結果を報告する．

第 2 章

先行研究

2.1 JO モデル

日本ネットワークセキュリティ協会 (JNSA) では，インシデント情報を集計し，被害人数，漏洩情報種別，漏洩原因，漏洩経路などを分析している．

専門家でなくても自らの潜在的リスクを算出できるように入力値を絞り，算定が容易となるような計算モデル [3] を次のように提案している．

$$
\begin{aligned}
&\text{損害賠償額} \\
&\quad = \text{基本情報価値 } [500] \\
&\qquad \times \text{機微情報度 } [10^{max(x)-1} + 5^{max(y)-1}] \\
&\qquad \times \text{本人特定容易度 } [1, 3, 6] \\
&\qquad \times \text{社会的責任度 } [1, 2] \\
&\qquad \times \text{事後対応評価 } [1, 2]
\end{aligned}
\tag{2.1}
$$

ここで，[ ] 内の値は，各値域である．機微情報度は図 2.1 で定められる 3 値を取る精神的レベル X と経済的レベル Y の 2 変数で与えられる．本人特定容易度は

$$
\text{本人特定容易度} = \begin{cases} 6 & \text{氏名 } \textit{and} \text{ 住所} \\ 3 & \text{氏名 } \textit{or}(\text{住所 } \textit{and} \text{ 電話番号}) \\ 1 & \text{その他} \end{cases}
$$

と定められている．

2.2 Romanosky モデル

[2] において，Romaosky は，米国の Advisen 社から 2004 年から 2015 年の間に記録された 12,000 件を超えるサイバーインシデントのデータセットを分析し，業界別及び時間の経過に伴うこれらのイベントのコストを次のようにモデル化した．

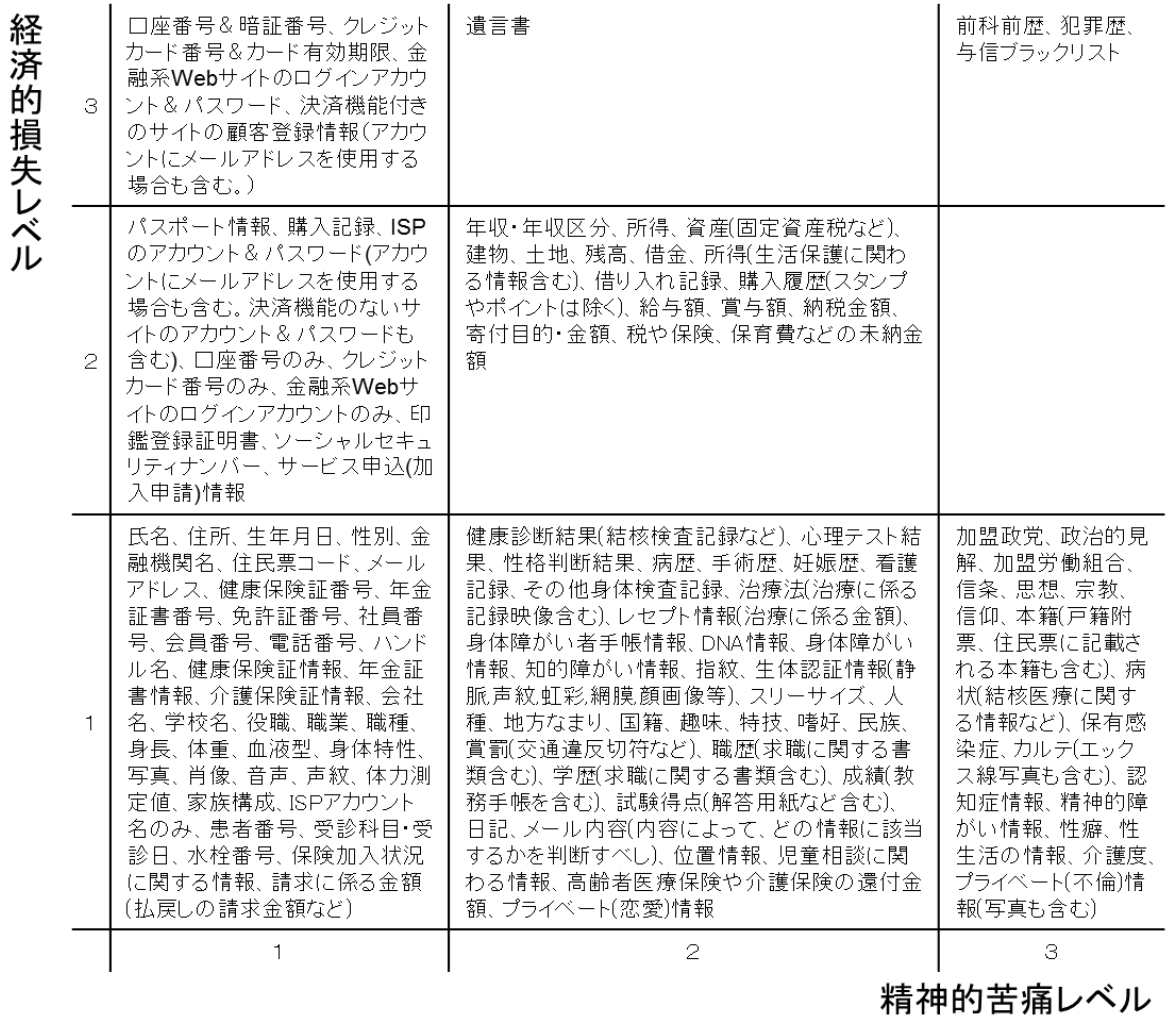

図 2.1　経済的レベル・精神的レベル EP 図 [3]

$$
\begin{aligned}
log(cost_{i,t}) = {} & \beta_0 + \beta_1 \cdot log(revenue_{i,t}) \\
& + \beta_2 \cdot log(records_{i,t}) \\
& + \beta_3 \cdot repeat_{i,t} + \beta_4 \cdot malicious_{i,t} \\
& + \beta_5 \cdot lawsuit_{i,t} + \alpha \cdot FirmType_{i,t} \\
& + \lambda_t + \rho_{ind} + \mu_{i,t}
\end{aligned}
\tag{2.2}
$$

ここで，各係数の値を表 2.1 に示す．i，t は t 年に企業 i が被ったインシデントを示し，$revenue$ は企業の収益，$records$ はインシデント被害にあった件数を示している．$repeat$，$malicious$，$lawsuit$ はブール値でそれぞれ企業が複数回インシデントに見舞われたかどうか，インシデントが悪意によって引き起こされたものかどうか，起訴されたかどうかを示し，肯定ならば 1 をとり，否定ならば 0 をとる．$FirmType$ はインシデントを受けた組織が政府機関，非営利企業，非公開企業，上場企業のいずれであるかを表すダミー変数である．λ_t，ρ_{ind}，$\mu_{i,t}$ はそれぞれ t 年のみを 1 とする年次ベクトル，業種を表すベクトル，エラー項である．

表 2.1　Romanosky の提案モデルの各係数 [2]

係数			Estimate	
定数	β_0		-3.858	*
$log(revenue_{i,t})$	β_1		0.133	**
$log(record_{i,t})$	β_2		0.294	***
$repeat_{i,t}$	β_3		-0.352	
$malicious_{i,t}$	β_4		-0.029	
$lawsuit_{i,t}$	β_5		0.044	
		Government	-1.032	
$FirmType_{i,t}$	α	Private	-1.032	
		Public	-0.065	

しかし，このモデルはアメリカの企業に対して最適化されており，日本の企業に対して同じように適用できない．

2.3　山田モデル

山田は，[1] において，2005 年から 2016 年までの JNSA データセットと QUICK Astra Manager より購入した本決算 (連結優先) データの個人情報漏洩インシデントが発生した年の会計情報を取得し，114 件分のインシデントデータを重回帰分析している．インシデントが発生した年の特別損失額 y を目的変数として回帰分析して，次のモデルを提案している．

$$\begin{aligned} log(y) = \beta_0 + \beta_1 \cdot log(x_1) + \beta_2 \cdot log(x_2) \\ + \beta_3 \cdot x_3 + \cdots + \beta_{16} \cdot x_{16} \end{aligned} \tag{2.3}$$

ここで，各係数の値と定義域を表 2.2 に示す．経済的ランク x_5 や本人特定容易度 x_6 などは JO モデルと同様である．

表 2.2　山田モデルの各係数

係数				Estimate	定義域
	定数			-3.9632	
	log(被害人数)	$log(x_1)$	β_1	0.0379	
	log(売上高)	$log(x_2)$	β_2	0.9904	
	故意	x_3	β_3	0.6261	0,1
	事後対応度	x_4	β_4	N/A	0,1
	経済的ランク	x_5	β_5	0.1590	1,2,3
	精神的ランク	x_6	β_6	0.0128	1,2,3
	本人特定容易度	x_7	β_7	0.2079	1,3,6
業種	不動産業，物品賃貸業	x_8	β_8	-0.0773	0,1
	建設業			-1.4450	
	情報通信業			-0.1350	
	林業			-0.4030	
	電気・ガス・熱供給・水道業			-0.9330	
	生活関連サービス業，娯楽業			-1.0040	
	卸売行，小売業			-0.4550	
	医療，福祉			-0.6319	
	宿泊業，飲食サービス業			-0.4607	
	製造業			-0.7577	
	教育，学習支援業			-0.0654	
	学術研究，専門・技術サービス業			-0.1173	
	金融業，保険業			-1.7570	
	運輸業，郵便業			-0.8893	
	氏名	x_9	β_9	-0.6231	0,1
	住所	x_{10}	β_{10}	-0.5169	0,1
	電話番号	x_{11}	β_{11}	-0.5337	0,1
	生年月日	x_{12}	β_{12}	-0.2348	0,1
	性別	x_{13}	β_{13}	0.2624	0,1
	職業	x_{14}	β_{14}	0.1453	0,1
	メールアドレス	x_{15}	β_{15}	-0.3845	0,1
	ID/PASS	x_{16}	β_{16}	-0.2810	0,1

第 3 章

サイトの開発

3.1　概要

サイトの開発には PHP を用いた．サイトは組織についての情報を入力するページと推定される損害額を出力するページの 2 つで構成されている．

3.2　入力項目

入力項目は図 3.2 にある被害人数 x_1，売上高 x_2 のような数値と故意 x_3，事後対応 x_4 のようブール値などの 18 項目である．

3.3　出力結果

JO モデル，Romanosky モデル，山田モデルの 3 種類のコスト評価値を表で整理した本サイトの実行結果を図 3.3 に示す．この例はベネッセホールディングスが 2014 年に起こしたインシデントである．

図 3.1　入力項目

図 3.2　出力結果 (ベネッセホールディングスの例)

第 4 章

評価

先行研究は 2018 年までのインシデントに基づいていたため，2021 年現在のインシデントデータに適用して有用性を調査する．また，ウェブサイトを実際に使ってもらい，その結果を示す．

4.1　データの概要

2021 年のインシデントの情報は，2000 年から情報漏洩による被害件数が 1000 件以上あった組織の情報をまとめているサイト [5] と各企業の決算情報をもとに本研究で収集した．本インシデントデータの統計量を表 4.1 に示す．

表 4.1　2021 年に起きたインシデントの統計量

期間	レコード数	企業数	属性数	平均被害人数
1 年間	143	108	22	530,917.29

インシデントデータ 109 件のうち売上高を公開している組織 47 件を用いる．

4.2　評価結果

2021 年のデータでの出力結果を各モデルと実際の被害金等の比較を表 4.2 に示す．説明変数の最も多い山田モデル誤差が少ない．Romanosky モデルは米国の企業を対象として分析を行っているため，日本の企業の推定誤差が大きい．

4.3　比較

2018 年のデータと 2021 のデータで行った推定値を比較する．2018 年は山田モデルの線形式と各インシデントの散布図を図 4.3 に示す．同様に 2021 年の山田モデル，JO モデル，Romanosky モデルを図 4.3，4.3，4.3 に示す．横軸は実際の損害額，縦軸は推定損害額である．線形式は実際の損害額に推定損害額が一致した場合のものである．比較する件数に差があるものの各モデルを見比べても大きな差が見られない．

表 4.2　2021 年の各モデルの推定損害額の比較

No	企業名	日付	被害人数	損害額	山田モデル	JO モデル	Romanosky モデル
1	柿安本店	2021/1/19	1293	805.7	1071.1	16.8	37.3
2	東京ガス	2021/2/1	10365	12498.9	16237.6	10.3	152.9
3	全日本空輸	2021/3/6	1000000	84252.2	10025.7	3000	216.5
4	メルカリ	2021/5/21	28889	92.5	11632.2	1476.4	289.2
5	日本航空	2021/3/5	920000	7877.0	6626.4	2760	189.5
6	JTB	2021/8/18	2525	4947.9	2493.4	7.5	80.2
7	宝ホールディングス	2021/3/19	4167	1591.8	3152.6	25	78.7
8	Coinbase	2021/9/28	6000	338.7	429.8	36	75.5
9	イオン銀行	2021/2/22	2062	151.9	437.6	160.8	62.2
10	東急コミュニティー	2021/3/29	5000	2169.1	10267.9	18.3	112.4
			平均	4285.3	3483.1	982.0	92.2
			最大	84252.2	22706.5	10270.5	350.6
			最小	0.05	15.7	2.3	15.4

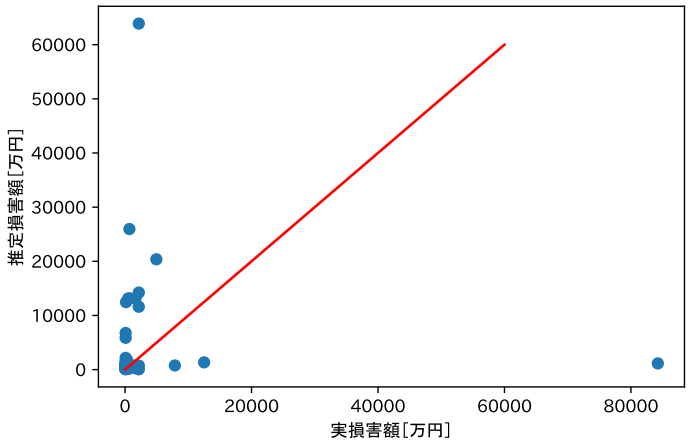

図 4.1　インシデントによる損害額　山田モデル (2021 年)

4.4　被験者実験

2021 年 12 月 13 日から 20 日までの 1 週間で，菊池研究室の 7 名に 2011 年 3 月 4 日に高島屋で起きた情報漏洩について損害金額の推定を試行してもらい，入力内容と出力結果の正しさと，その際にかかった時間を調査する．

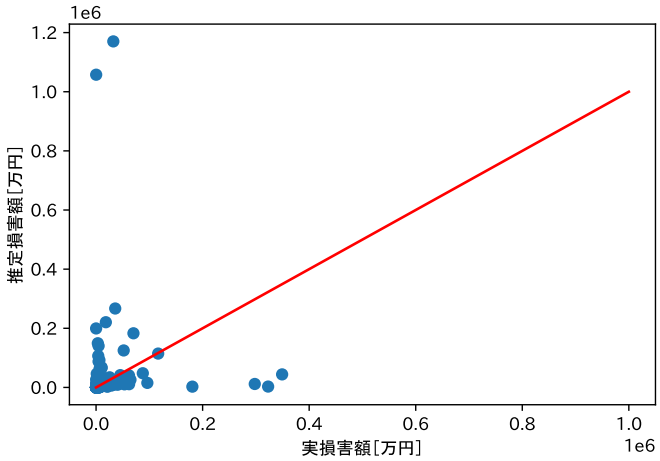

図 4.2　インシデントによる損害額　山田モデル (2018 年)

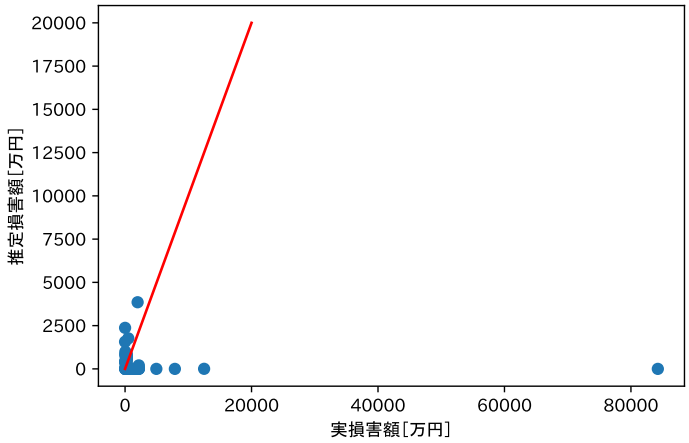

図 4.3　インシデントによる損害額　JO モデル (2021 年)

4.5　実験結果

出力結果の山田モデルについての誤差と実験時間を表 4.3 に示す．誤差の原因は売上高の検索で誤った場所を見ているためであることが分かった．

表 4.3　被験者実験の結果

	平均誤差 [万円]	平均時間
平均	29907.5	11 分 28 秒
最大	1,325,890.7	22 分
最小	0	6 分 36 秒

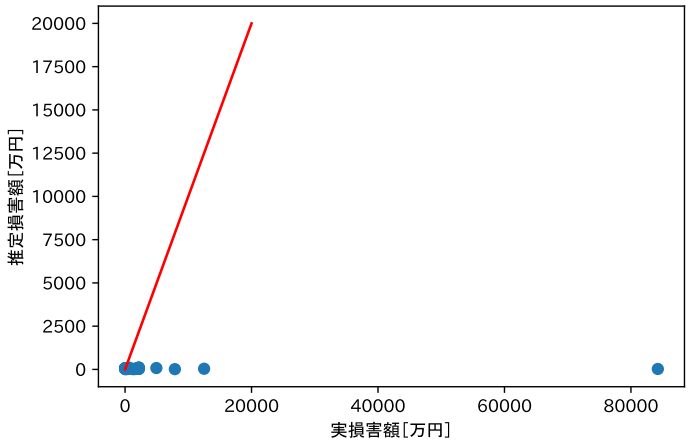

図 4.4　インシデントによる損害額　Romanosky モデル (2021 年)

また，被験者からの意見として

- どの決算データのどこを見ればいいのかわからない
- 社会的責任度の知名度の基準が分からない

などがあった.

第 5 章

おわりに

本研究では PHP を用いて一般の人でも組織のインシデント損害金を推定できるようなサイトの開発を行った．推定する目的の違う各モデルを一度に確認できるようになり，専門知識のない一般の人でも簡単にインシデント被害金額を推定できるウェブサイトの開発ができた．被験者実験で出力結果を誤って読むようなサイトになっていることが分かった．インターフェースの改善を今後の課題とする．

第 6 章

謝辞

本研究を行うにあたり，多くの方より御指導いただきました．特に明治大学総合数理学部先端メディアサイエンス学科，菊池浩明教授に深く感謝申し上げます．また, メンターとしてご指導いただいた堀込光さん, 研究室の皆様に深く感謝の意を表するとともに, 謝辞とさせていただきます.

参考文献

[1] 山田道洋，菊池浩明， 松山直樹，乾孝治，”個人情報漏洩の損害額の新しい数理モデルの提案”，第 80 回 CSEC 研究報告会, pp.2–3，2018.

[2] Romanosky, S.: ”Examining the costs and causes of cyberincidents”, Journal of Cybersecurity, Vol.2, No.2, pp.121–135，2016.

[3] 情報セキュリティインシデントに関する調査報告書別紙 (https://www.jnsa.org/result/incident/data/2016incident_survey_attachment_ver1.0.pdf，2021 年 11 月参照).

[4] 不正アクセスによる会員様情報流出に関するお詫びとお知らせ (https://www.net-marketing.co.jp/news/5873/，2021 年 5 月参照).

[5] 個 人 情 報 漏 洩 事 件・被 害 事 例 一 覧 (https://cybersecurity-jp.com/leakage-of-personal-information，2021 年 11 月参照).

付録 A

疑似人流データを用いた感染確率の推定

A.1　はじめに

新型ウイルスは流入・発生した段階では医療体制は整っていない可能性があり，2019 年 12 月に流行した新型コロナウイルスは急速に感染を拡大し，2020 年 12 月 18 日現在，日本では 2661 人の死者を出している [2]. さらには緊急事態宣言の発令によって経済を滞らせている．感染拡大の防止策としてテレワークや時差出勤などがあるが，各策の効果が不明であり，不安が蔓延している．

そこで，本研究では，どれくらいの人数が感染している時どのような行動をすると感染症は収束するのか，厚生労働省からリリースされた新型コロナウイルス接触確認アプリ (COCOA) をインストールしたことによる効果を明らかにすることを目的とする．感染者の推移をシミュレーションするために，大規模の移動履歴を表したナイトレイ社 [1] 疑似人流データを用いて感染確率分布の推定を試みる．

A.2　感染推定

A.2.1　実験環境

本研究では，疑似人流データは株式会社ナイトレイが提供している関東圏のデータ [1] を使用する． 本データは 2013 年 7 月 1 日における関東圏の 6432 人分の 0 時から 24 時までの 5 分毎の緯度経度情報であり，1 日の時系列データである．データの例を表 A.1 に示す．属性の中に状態を示すものがあり，位置情報が変化するときは MOVE，位置情報が変化しない時間が始まる場合は STAY となっている．本実験では，データの補完は R，感染確率の推定は Python で行った．

表 A.1　疑似人流データ (一部)

ユーザ ID	日付・時刻	緯度	経度	状態
1013	2013/7/1 0:00	35.717	139.899	STAY
1013	2013/7/1 8:15	35.676	139.787	STAY
1013	2013/7/1 9:45	35.676	139.899	MOVE
1013	2013/7/1 9:50	35.676	139.899	MOVE

A.2.2　実験方法

6432 人から初期の感染確率 p に従ってランダムに感染者を決める．未感染者が感染者と 3 メートル以内で 15 分以上接近すると 50 %で感染する．p を 5 %から 50 %まで 5 %毎に変更して結果を比較する．

A.2.3　データの補完

本疑似人流データは表 A.1 で示したように，0 時から 24 時まで全ての 5 分毎の位置情報を格納しているわけではない．そのため，位置情報がない時間は直前の位置情報を補完する．補完後のデータを表 A.2 に示す．*印が補完されたレコードである．

表 A.2　疑似人流データ補完後

ユーザ ID	日付・時刻	緯度	経度	状態
1013	2013/7/1 0:00	35.717	139.899	STAY
1013	2013/7/1 0:05	35.717	139.899	STAY*
⋮	⋮	⋮	⋮	⋮
1013	2013/7/1 8:10	35.717	139.899	STAY*
1013	2013/7/1 8:15	35.676	139.787	STAY
1013	2013/7/1 8:20	35.676	139.787	STAY*

A.2.4　予備実験

新型コロナウイルス接触確認アプリ (COCOA) の仕様によると，1 メートル以内で 15 分以上の接近を接触として検知する．この基準に従って新宿に 8 時間以上滞在していた 157 人分の位置情報データで接触回数を調べたところ，接触回数がほとんど確認されなかった．そこで，本研究では範囲を少し広く定めて，3 メートル以内で 15 分以上接近を接触として 6432 人分のデータでの評価を行う．

A.2.5　実験結果

接触回数と感染確率の結果は表 A.3 に示すようになった．接触回数は 30 分毎の延べ回数である．

接触の頻度分布を図 A.1 に示す．接触は，8 時から 9 時の通勤・通学の時間帯，12 時の昼食の時間帯，18 時ごろの退勤の時間帯にピークがある．最も接触回数の多くなるのは 18 時 30 分であった．

初期の感染確率が 5 %の時は 24 時間で 0.5 %の増加率 p'/p と推定された．初期の感染確率に比例して増加率も増えている．

A.3　可視化

補完後の疑似人流データを用いて，0 時から 24 時まで 5 分毎で位置をプロットする可視化プログラムを Processing で開発した．0 時と 12 時の位置分布を各々図 A.2 と A.3 に示す．昼夜の人の密度の差から，多くの人が電車などの公共交通機関を使っていることがわかる．

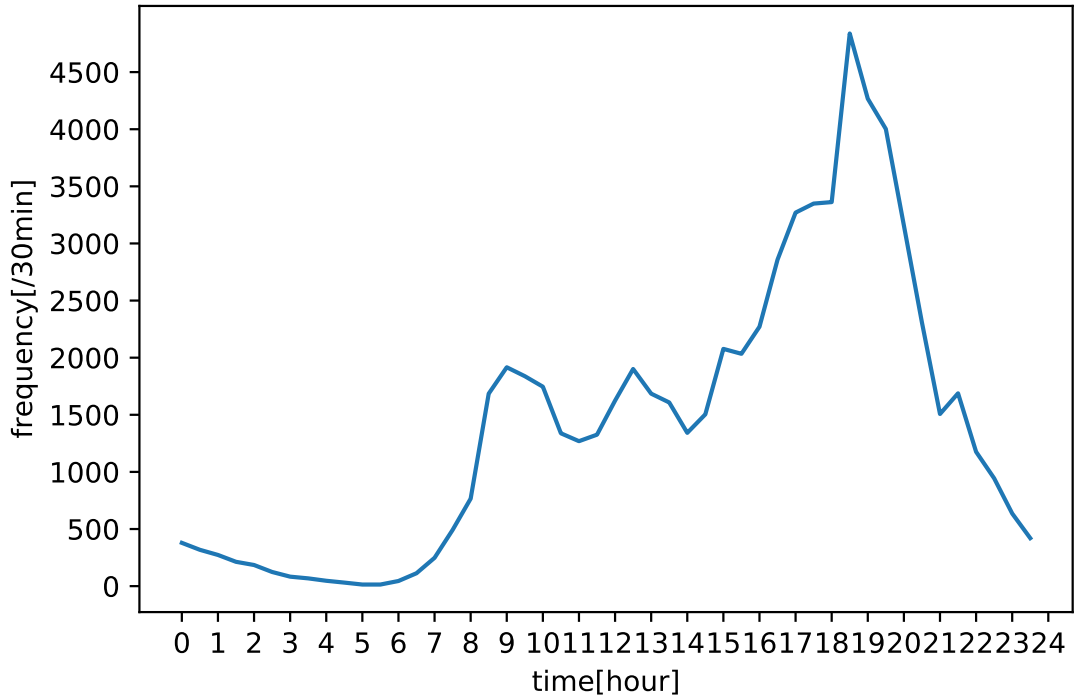

図 A.1　30 分毎の延べ接触回数の分布

表 A.3　初期確率 p による感染確率

感染確率 p[%]	24 時間後の感染確率 p'[%]	増加率
5	5.5	0.1
10	12.1	0.21
15	17.9	0.19
20	22.7	0.14
25	27.7	0.11
30	33.6	0.12
35	37.3	0.07
40	50.0	0.25
45	55.2	0.23
50	58.1	0.16

A.4　おわりに

本稿は，感染対策の効果や新型コロナウイルス接触確認アプリ (COCOA) の効果を求める目的のための予備実験を報告した．目的達成のためには，外出自粛状態になった場合に疑似人流データをどのように操作するのか，COCOA からの通知が来たときの操作などといった課題を解決する必要がある．

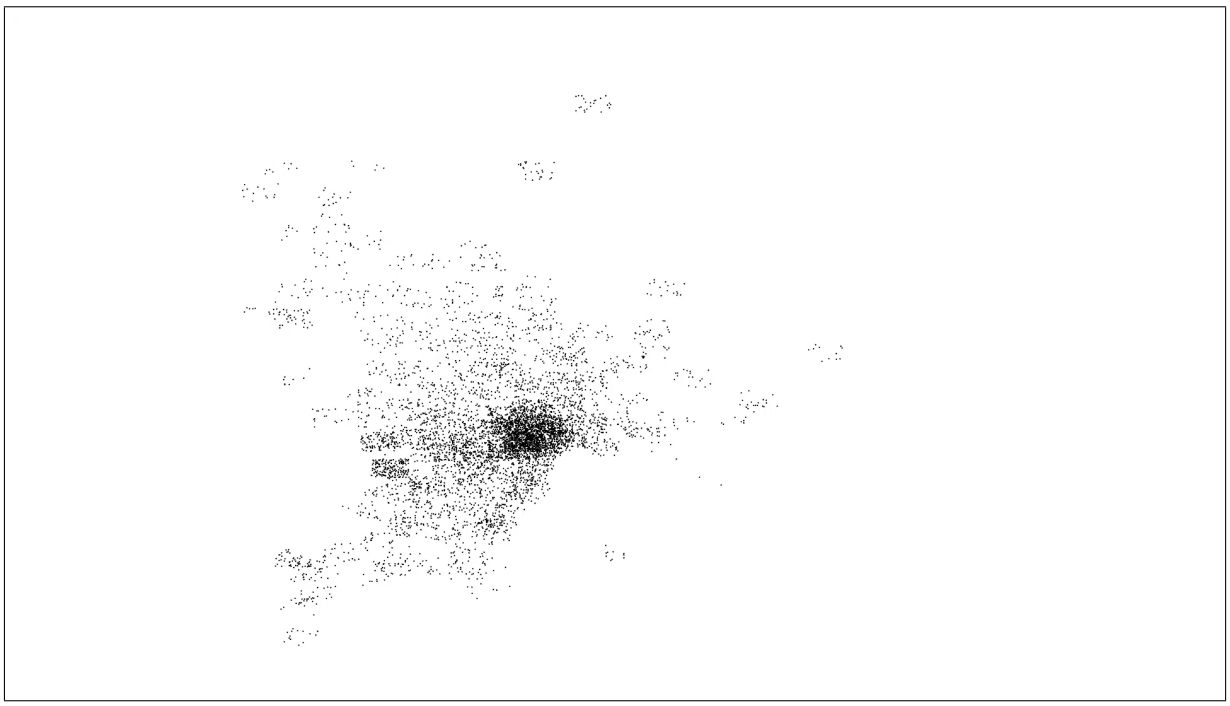

図 A.2　2013 年 7 月 1 日 0 時の位置情報

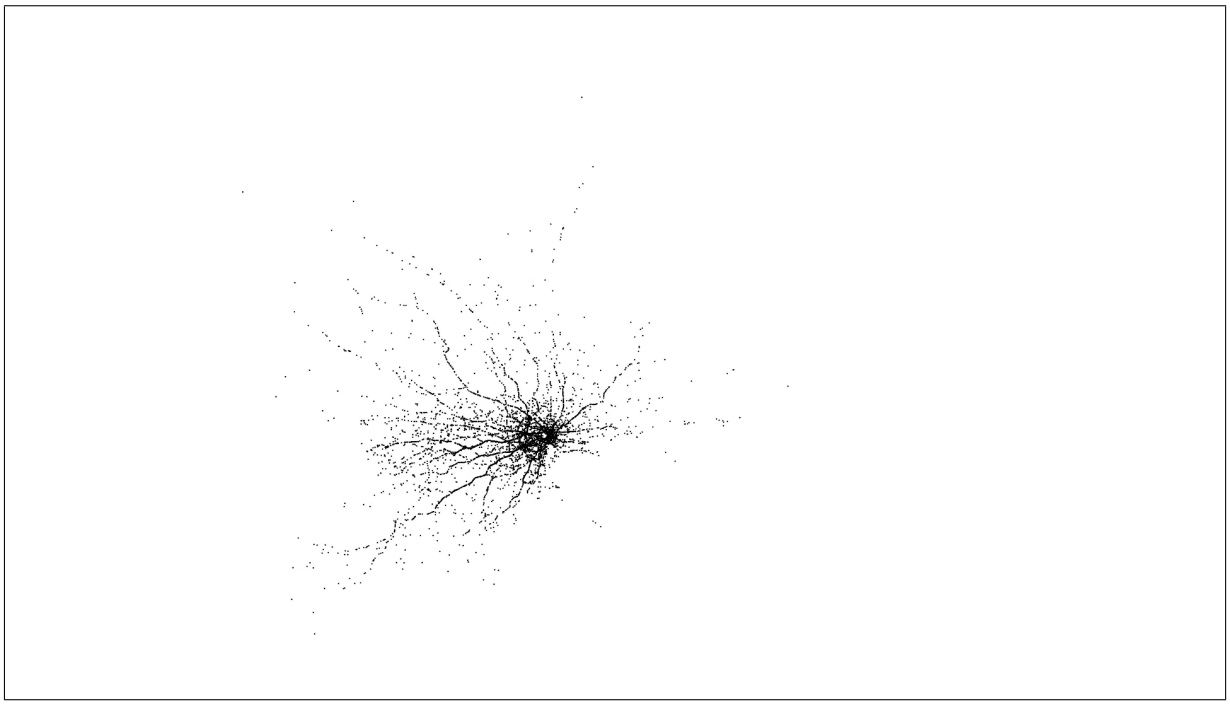

図 A.3　2013 年 7 月 1 日 12 時の位置情報

参考文献

[1] 疑似人流データ，株式会社ナイトレイ (https://nightley.jp/archives/1954/，2020 年 8 月参照)

[2] 日本国内の感染者数（NHK まとめ）(https://www3.nhk.or.jp/news/special/coronavirus/data-all/，2020 年 12 月参照)

[3] 新型コロナウイルス接触確認アプリ (https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/covid19_qa_kanrenkigyou_00009.html#cocoa1，2020 年 12 月参照)

[4] 倉橋 節也，新型コロナウイルス (COVID-19) における感染予防の推定，人工知能学会論文誌，p.1(https://www.jstage.jst.go.jp/article/tjsai/35/3/35_D-K28/_pdf , 2020 年 12 月参照)