修士論文

リソース・モデル・エリアを考慮した
インフルエンザ流行予測

村山 太一

2019 年 3 月 12 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

村山 太一

審査委員：

松本 裕治 教授	（主指導教員）
中村 哲 教授	（副指導教員）
荒牧 英治 特任准教授	（副指導教員）
鈴木 優 特任准教授	（副指導教員）

リソース・モデル・エリアを考慮した インフルエンザ流行予測*

村山 太一

内容梗概

インフルエンザは世界的にも多くの死者をもたらす感染症の 1 つであり，公衆衛生機関に多くの負担をもたらす．そのことからも，インフルエンザの流行予測を行うことは重要な課題である．既存の研究では，インフルエンザの流行予測に必要なりソースとして User-generated contents や過去のインフルエンザ罹患者数情報が用いられ，様々なモデルが適用されてきた．本論文では，予測のためのリソース，新たなモデル，そしてあまり行われてこなかったエリアの 3 つに焦点をあてた研究を行なう．リソースに関する実験では，これまで用いられてきた多くのリソースを用いてインフルエンザ流行予測を行い，リソースごとの性質や効果の検証を行なう．次に，1 つ目の実験をもとに，User-generated contents と過去の罹患者数情報を別々に学習を行う新たなモデル Two-stage Model を提案する．最後に，時空間モデルと人流データを用いて，地域ごとのインフルエンザ流行予測を行なう．

キーワード

時系列処理，公衆衛生，インフルエンザ，User-generated content，Infodemiology，機械学習，予測

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文,
2019 年 3 月 12 日.

Influenza epidemics prediction considering resource, model and area*

Taichi Murayama

Abstract

Influenza, an infectious disease, causes many deaths and work load for medical institutions worldwide. Therefore, predicting influenza patient volume during epidemic periods is an important task. Data of two kinds are often used for the prediction: user-generated content (UGC) and historical Influenza-like illness (ILI) data, and various models are developed in many researches. I conduct three experiments focusing on resources, models and areas, to which little attention has been paid. About resources, I predict the influenza volume with many resources, and validate the effectiveness and nature of each resource. Secondly, I propose new model Two-stage Model contrived how to use resources, based on first experiment. Finally, Spatio-temporal model with commuting data predicts the influenza volume by region.

Keywords:

time-series analysis, public health, Influenza, User-generated content, Infodemiology, Machine Learning, prediction

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, March 12, 2019.

目次

図目次

第 1 章　はじめに

1.1　背景

感染症流行予測は公衆衛生機関にとって重要である．近年は，感染症流行サーベイランスや予測のための重要なリソースとして，インターネット上の情報を活用する，Infodemiology [4] と呼ばれる分野も注目を集めている．その中でも，検索履歴や SNS 投稿などの Online User-Generated Contents (UGC) は，人々のオンライン上での行動だけでなく実生活の行動についても多くの情報を反映しており，感染症流行予測に対しても流行を早期に感知できる [1]．代表例として，検索クエリを用いたインフルエンザ予測を行う Google Flu Trends [3] が挙げられ，検索クエリやソーシャルメディアへの投稿などの情報が重要なリソースの 1 つとなっている．

また，近年の人工知能技術の発展により感染症流行予測のためのモデルにおいても，ARIMA モデル [30] や Random Forest [17]，LSTM [5] などの深層学習を用いたモデルが適用されてきている．

本論文では，感染症の中でも世界的に流行が生じるインフルエンザについて扱う．インフルエンザは，毎年 29 万 – 65 万人が死亡することで知られており，流行によって労働者の欠勤や医療従事者や医療関係者の負担の増大などによって経済的損失を生み出す．このことからも，1 週間以上先のインフルエンザの流行予測は，医療機関の人材配分や管理の面からも重要である．

日本のインフルエンザは季節性の特徴が見られるが，2009 年における豚インフルエンザ流行など季節性ではない突発的な流行もあり予測を難しくする要因の一つとなっている．図 1.1 に，日本における最近 10 年の定点あたりインフルエンザ罹患者数を示す．

1.2　本研究の概要

これらの背景を踏まえて，本論文ではリソース，モデル，そして人流を考慮したエリアの 3 つに焦点を当て，インフルエンザ流行予測実験を行う．

はじめに，リソースに関する研究を行う．インフルエンザ流行予測のためのリ

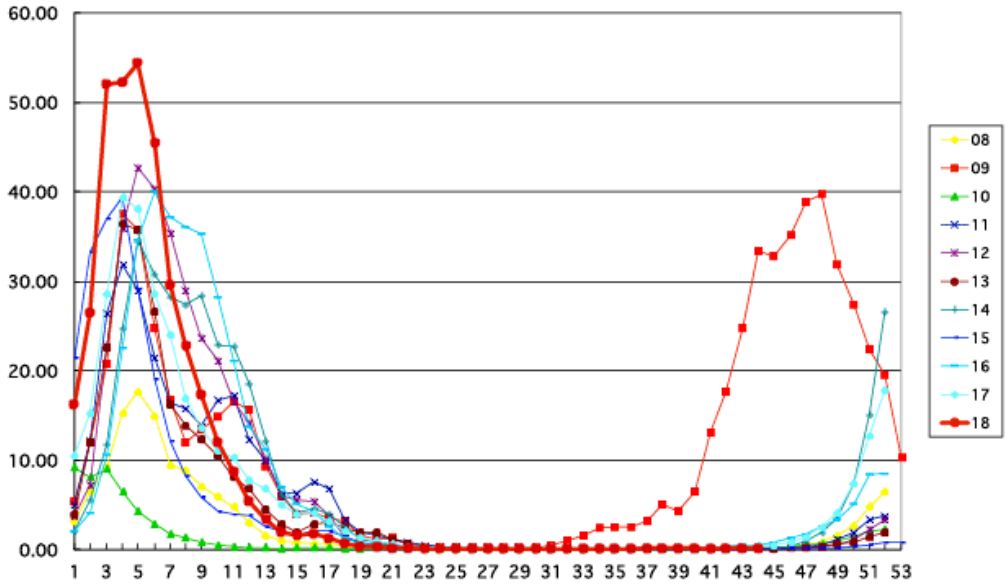


図 1.1　日本の 2008 - 2018 年における定点あたりのインフルエンザ罹患者の割合（出典：国立感染症研究所 (NIID)）．日本は 12–2 月の寒気に流行が生じる季節性の特徴を持っているが，2009 年の新型インフルエンザの流行などの季節性と異なる流行が生じる場合もある．

ソースは多様化していることから，どのリソースを選択するべきか，もしくはどのリソースを組み合わせてインフルエンザ流行予測のモデル構築を行えばよいのかという問いが生じる．多くの研究が対象としている国，時期，リソースが異なることからも，リソース間の比較や各リソースの性質の検討を他の研究を参照しまとめて行うのは困難である．そこで，これまで用いられてきたリソースを用いてインフルエンザの予測を行い，リソースごとの性質や効果の検証を行う．

次に，最初の実験で得られた洞察を元に，リソースの使い方を工夫した新たなインフルエンザ流行予測モデルを提案する．具体的には，インフルエンザ流行予測に対して Online User-Generated Contents (UGC) とインフルエンザ罹患者データの 2 つのリソースが主に用いられているが，これらの利点を考慮した新たなモデル Two-stage Model を作成する．このモデルの特徴は，季節性など時期列固有部分を regular trend，突然の予期できない変動部分を irregular trend とみなし，それらを別々のモデルで学習する点である．

最後にエリアに関する研究として，地域間の関係を考慮してインフルエンザ流行

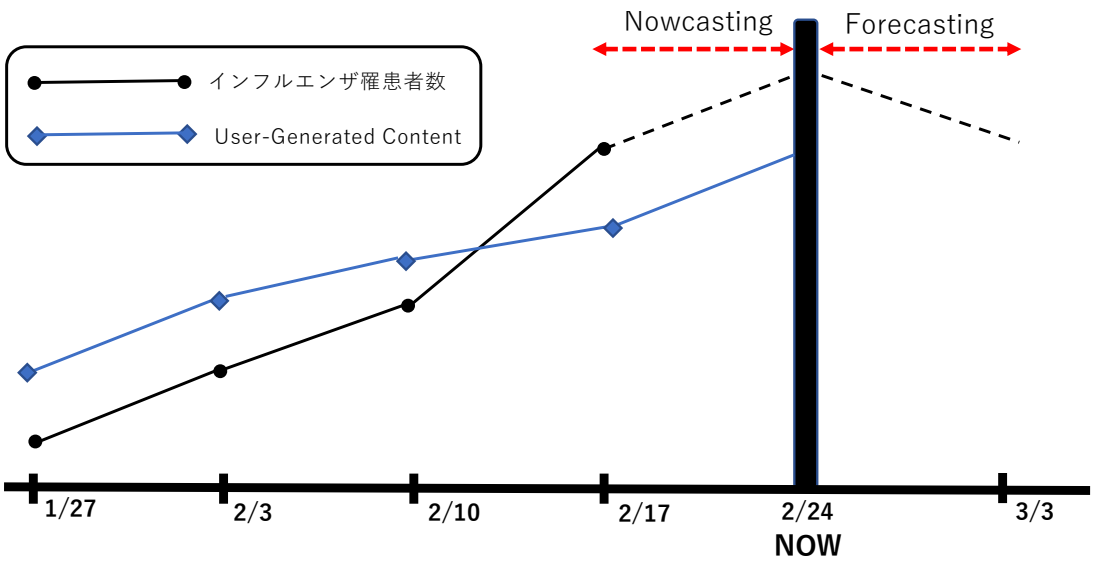


図 1.2　本論文におけるインフルエンザ**予測**の概要図．黒線はインフルエンザ罹患者数，青線は UGC における検索（投稿）数を指す

予測モデルの作成を行う．インフルエンザ流行予測では国を単位として行われるのが一般的であるが，実用においては，より詳細な地域ごとの流行予測が求められる．このことから，通勤・通学データと時空間モデルを用いて，日本の各都道府県の流行予測を行う．

本論文の目的である，本研究におけるインフルエンザの「予測」について説明する．一般的に，インフルエンザの罹患者データは 1 週間程度の遅れを伴って提供されるという特徴を持っていることから，罹患者データ提供時点から 1 週間の予測は “Nowcasting” と呼ばれ，2 週間先以降の予測については “Forecasting” と呼ばれる．本論文においても，インフルエンザ**予測**とは，特に断りのない場合，“Forecasting”，つまり 2 週間先の予測を指す．また，インフルエンザ罹患者数の報告の各定点の 1 週間における合計値となるため，予測も 1 週間の合計値を求める．その際に利用するリソースは実用上の観点から現時点で取得可能なものに限る．

図 1.2 に概要図を示す．例えば，2/24 を現時点としたとき，UGC においては 2/24 までのデータを利用できるが，インフルエンザ罹患者数のデータは 2/17 までのデータしか利用できない．ここでの “Nowcasting” とは，これらのデータを用いて 2/18–2/24 におけるインフルエンザ罹患者数の合計値を推定することを指

す．本論文でのインフルエンザ予測とは，特に断りのない限り，“Forecasting”，つまり 2 週間先の予測を指す．つまり，これらのデータを用いて “Forecasting” の 2/25–3/3 のインフルエンザ罹患者数の推定を “予測” と述べる．

1.3　本論文の構成

本論文の構成は以下のとおりである．まず，2 章でインフルエンザ流行予測に関する関連研究について述べる．3 章ではリソースに焦点を当てた「インフルエンザ流行予測における複数リソースの検証」，4 章ではインフルエンザ流行予測のための新たなモデルを提案する「レギュラー部分とイレギュラー部分に分割したインフルエンザ予測」，5 章ではエリアに着目した「人流を考慮した地域ごとのインフルエンザ流行予測」，そして最後に 6 章で本論文を通してのまとめや展望について述べる．

第 2 章　関連研究

本論文のテーマとする「リソース」,「モデル」,「エリア」の 3 つ観点から，インフルエンザ流行予測の関連研究について述べる．

2.1　リソース

感染症を予測するためにどのようなリソースを利用するかという問いは Infodemiology [4] の分野で重要なトピックである．予測の際に用いられるリソースとして UGC が多いと前章で述べたが，その中でも特に検索クエリ [3, 6, 7, 8, 10] や SNS(Twitter) [10, 11, 12, 13, 14] が用いられることが多い．この 2 つのリソース，検索クエリや Twitter のどちらが有用かに関する議論は多く行われている [7, 14, 15]．検索クエリや Twitter 以外の UGC として，Wikipedia [7] やインフルエンザ専用の報告アプリ [18] などを用いた研究が挙げられる．流行予測のために用いられれる UGC 以外のリソースとして，症状に対する看護師からのアドバイスを貰える Telephone triage [19] や医薬品販売 [20, 21]，遺伝子配列 [22, 23]，天気データ [16, 17] などが存在する．また，インフルエンザ流行予測は情報学的には時系列予測にあたることから，過去のインフルエンザ罹患者数をリソースとする研究も多く見られる [8, 24].

どのようなリソースを用いるかと同時に，どのようなクエリを用いるかという問いも重要である．クエリ選択ではインフルエンザ罹患者数と検索するとの相関を元に選択することが一般的であるが，他にも Google Correlate を用いた手法 [25] や Word Embedding を用いた手法 [26]，などが考案されている．Twitter を用いたものとして，罹患者による投稿かどうかの分類を行い，罹患者による投稿と判断されればそれを予測のリソースとする手法 [27, 28] なども提案されている．

2.2　モデル

インフルエンザ流行予測では線形回帰を用いられることが多いが [3, 11]，よりインフルエンザ流行予測に適したモデルを探索するために様々なモデルが適

用・考案されている．時系列分析ベースの手法として，Box-Jenkins の 1 種である Auto-Regression Integrated Moving Average (ARIMA) [30] や Generalized Auto-Regressive Moving Average (GARMA) [34] が用いられている．機械学習ベースの手法として，Random Forest [30, 17]，Long Short Term Memory (LSTM) [24]，マルチタスクガウス過程 [29] などが用いられている．また，数理モデルに基づいたものとして，IDEA モデルなどの感染症モデル [31, 32] や Bayesian Model Averaging [6, 33] などが挙げられる．

本論文では，Random Forest を 4 章のベースラインに LSTM を 5 章のベースラインとして用いる．

2.3 エリア

地域ごとのインフルエンザ流行予測を行った研究として，マルチタスクガウス過程回帰を用いアメリカの HHS Regions を対象に流行予測をおこなったもの [29] などが挙げられる．本研究に関連する地域間の関係を考慮し地域ごとのインフルエンザ流行予測を行うものとして，地域間の距離に基づいたカーネルを用いて空間的依存性を捉えたもの [47] や CNN 機構を用いて求めたい地域の近傍を畳み込んだもの [48] などが挙げられる．

第 3 章　リソース：インフルエンザ流行予測における複数リソースの有用性の検証

3.1　本章の概要

関連研究でも述べたように，インフルエンザ流行予測のためのリソースは多様化している．このことからも，どのリソースを選択する，もしくはどのリソースを組み合わせてインフルエンザ流行予測のモデル構築を行えばよいのかという問いを明らかにすることが重要である．しかしながら，UGC を用いた多くの流行予測研究は 1 つないし 2 つのリソースを同時に用いるのに留まっている．3 つ以上の UGC を同時に用いた例として，Twitter，Google Trends，インフルエンザレポートを用いた Santillana et al. [10] が挙げられるのみである．また，多くの研究が対象としている国，時期，リソースが異なることからも，リソース間の比較や各リソースの性質の検討を他の研究を参照しまとめて行うのは困難である．

上記の課題から，本章の研究では 5 つのリソース (1) 過去の罹患者数 (2) 検索クエリ (3) Twitter の投稿 (4) ショッピングクエリ (5) Q&A サービスのクエリを用いて線形・非線形回帰のモデルによる日本のインフルエンザ流行予測モデルの作成，そして各リソースの有用性の検討を行う．

3.2　手法

3.2.1　データセット

本実験では，日本の 2013/10/12 - 2018/05/21 の期間のインフルエンザ流行予測を 5 つのリソースを用いて行う．この 5 つのリソースとして，過去の罹患者数を提供する国立感染症研究所 (NIID)，検索エンジン (Yahoo! Japan)，SNS (Twitter)，オンラインショッピング（Yahoo! ショッピング），Q&A サービス（Yahoo! 知恵袋）からデータを取得する．

国立感染症研究所 (NIID)

本実験では，国立感染症研究所 (NIID) *が提供しているデータを，過去のインフルエンザ罹患者数と正解データとして用いる．感染症流行把握のために集計しており，その結果を感染症発生動向調査週報 (IDWR)†を通して公開している．1 週間分の定点ごとの合計罹患者数が報告され，その報告は集計のために約 1 週間程度の遅れを伴って提供される．本実験では，インフルエンザ罹患者の予測のために，予測したい時点から 2 週間前から 53 週間前，つまり 52 個の集計結果を特徴として用いる．

検索エンジン

日本で多くのユーザに用いられている Yahoo! Japan‡の検索エンジンのクエリをリソースとして利用する．クエリ選択は，2013/10/02 - 2018/05/21 の「インフル」もしくは「インフルエンザ」を含んだツイートのうち，高頻出語上位 100 語と tf-idf の値が上位 50 語，つまりインフルエンザに罹患したユーザーやその家族が送信したと考えられる陽性ツイートのみに特徴的に出現する語を候補語として抽出する．次に，2013/10/12 - 2015/05/24 の期間における候補語の検索頻度と実際のインフル罹患者数との相関係数が 0.70 以上の 13 語の検索量を特徴量として選択する．

SNS

SNS の 1 つである Twitter§に投稿されているツイートをリソースとして用いる．ツイートを取得するためのクエリの選択は検索エンジンの方法と同様に，「インフル」もしくは「インフルエンザ」を含んだツイートのうち候補語が含まれたツイートを抽出する．そして，2013/10/12 - 2015/05/24 の期間での候補ツイート数と実際のインフル罹患者数との相関係数が 0.75 以上の 18 語が 1 つでも含まれたツイート数を特徴量として選択する．

---

*https://www.niid.go.jp/niid.html
†https://www.niid.go.jp/niid/ja/idwr.html
‡https://www.yahoo.co.jp/
§https://twitter.com/

オンラインショッピング

Yahoo! ショッピング[¶]をオンラインショッピングのサービスとして用いて，そのサイトにおける特定の検索クエリの検索数を特徴とする．ショッピングサイトでは商品名が検索されることから，我々が恣意的に選択した「マスク」などのインフルエンザと関連する 10 語をクエリとし，その検索量を特徴として選択する．

Q&A サービス

Yahoo! 知恵袋[‖]を Q&A サービスとして用いて，そのサイトにおける特定の検索クエリの検索数を特徴とする．オンラインショッピングと同様に，我々が恣意的に選択した「インフルエンザ A 型」などのインフルエンザと関連する 9 語をクエリとし，その検索量を特徴として選択する．

リソースの例や概要を表 3.1 と図 3.1 に，5 つのリソースの正規化したデータ量を図 3.2 に示す．

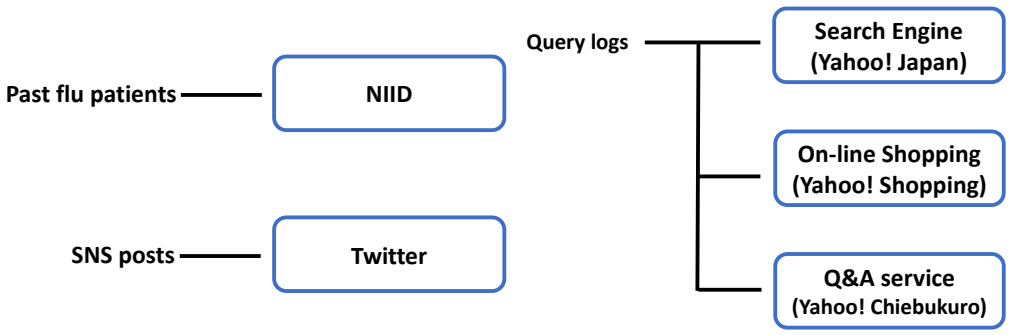


図 3.1　リソースの概要図

[¶]https://shopping.yahoo.co.jp/
[‖]https://chiebukuro.yahoo.co.jp/

表 3.1　リソースの詳細

	クエリ選択方法	クエリ例	特徴数
(a) 過去の罹患者数	-	-	52
(b) 検索エンジン (Yahoo! Japan)	相関係数 0.70 以上のクエリを選択	「高熱」 「いんふる」	13
(c) SNS (Twitter)	相関係数 0.75 以上のクエリを選択	「兄」 「辛い」	18
(d) オンラインショッピング (Yahoo! Shopping)	インフルエンザに関連する語を任意に選択	「空気清浄機」 「マスク」	10
(e) Q&A サービス (Yahoo! 知恵袋)	インフルエンザに関連する語を任意に選択	「インフル」 「A 型インフルエンザ」	9

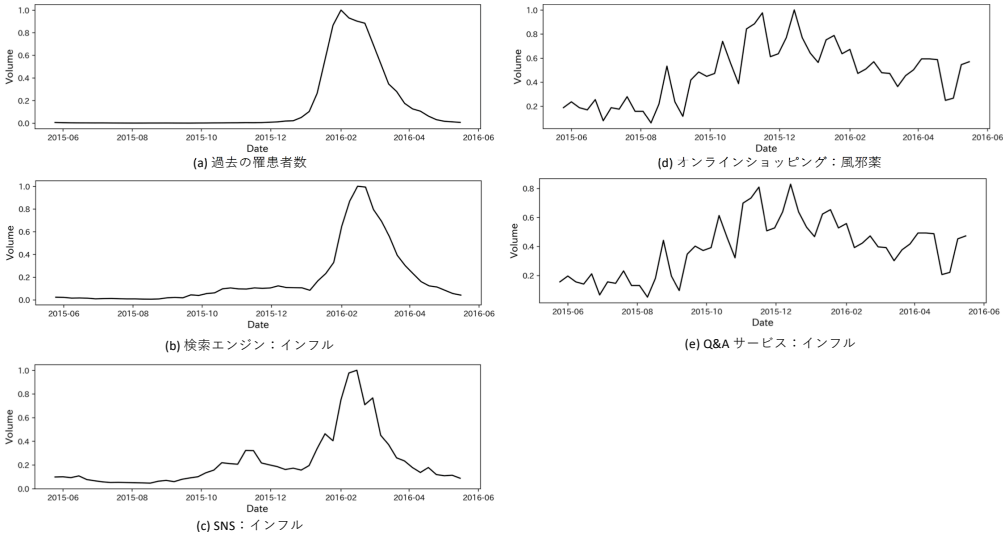


図 3.2　2015/06 - 2016/06 における各リソースのデータ量：(a) インフルエンザ罹患者数 (b) 検索エンジン クエリ「インフル」 (c) SNS クエリ「インフル」 (d) オンラインショッピング クエリ「風邪薬」 (e) Q&A サービス クエリ「インフル」

3.2.2　モデル

上記のリソースを用いてインフルエンザを予測するために，Lasso 回帰，Huber 回帰，Support Vector Machine Regression with Linear kernel (SVR)，Random Forest の 4 つの機械学習アルゴリズムを適用する．

Lasso 回帰

Lasso 回帰とは，線形回帰モデルのパラメータ推定の際 L1 正則化 (Lasso) を加えることで，変数選択と正則化を同時に行い，生成される統計モデルの予測精度と解釈可能性を上げる手法である．スパースな解が得られ，多くの変数の中から利用する変数を絞りこむことができるため，多くの場面で利用されている．Lasso 回帰は以下の式で定義される．

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{3.2.1}$$

$$\beta = \underset{\beta}{argmin}||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1 \tag{3.2.2}$$

$\mathbf{y}$ は観測ベクトル，$\mathbf{X}$ は特徴量ベクトル，β は回帰変数ベクトル，ϵ は誤差ベクトル，λ は L1 係数を表している．式 (3.2.1) は重回帰モデル式となっている．式 (3.2.2) の $\lambda||\beta||_1$ が L1 正則化の部分となっており，最小二乗法と L1 正則化を用いて最適な β を求める．

Huber 回帰

Huber 回帰 [36] は他の観測値パターンに従わない外れ値に対して頑丈な回帰を行うロバスト回帰の 1 つである．通常の損失関数で用いられる最小二乗法ではなく，式 (3.2.4) で定義される Huber 損失関数を用いる．

$$a = \frac{||\mathbf{y} - \mathbf{X}\beta||_1}{\sigma} \tag{3.2.3}$$

$$L_\delta(a) = \begin{cases} a^2 & (|a| \leq \delta) \\ 2|a| - 1 & (|a| > \delta) \end{cases} \tag{3.2.4}$$

Huber 損失のパラメータ δ には 1 が用いられることが多い．最小二乗法と比較して，Huber 損失は外れ値に対する損失の値が小さくなるという特徴をもつ．

Support Vector Machine Regression with Linear kernel (SVR)

Support Vector Machine Regression [37]（以降，SVR）は，「マージン最大化」を基本アイデアとする Support Vector Machine を回帰に拡張した手法である．カー

ネル法を用いることで，特徴空間での変数を陽に扱わないという特徴を持つ．SVR の回帰式を以下に示す．

$$K(x^{(j)}, x^{(i)}) = \phi(x^{(i)})\phi(x^{(j)})^T \tag{3.2.5}$$

$$f(x^{(j)}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x^{(j)}, x^{(i)}) + c \tag{3.2.6}$$

式 (3.2.5) の $\phi(x^{(i)})$ は $x^{(i)}$ の特徴ベクトルを表し，$K(\cdot)$ はカーネル関数を表す．式 (3.2.6) の α_i, α_i^* はラグランジェ定数で SVR の回帰式となる．本実験では，SVR のカーネルとして Linear カーネルを用いる．

Random Forest (RF)

Random Forest [38] は，相関の低い複数の決定木を用い，各決定木を弱学習木として扱うアンサンブル学習アルゴリズムの 1 つである．ノイズを多く含むが近似的にバイアスがないモデルの平均をとるという，バギングの考え方を取り入れており，過学習を防ぐなどの特徴が挙げられる．

3.3 実験

3.3.1 設定

5 つのリソースと 4 つのモデルを用いてインフルエンザの予測を行う．インフルエンザ罹患者データは 52 週間分，他のリソースは 1 週間前の検索数あるいは投稿数を特徴として用いる（3.2.1 節参照）．予測性能を評価するために，日本のインフルエンザは 3 シーズン (2015/05/23 – 2016/05/22, 2016/05/23–2017/05/21, 2017/05/22–2018/05/21) をテスト期間として設定する．訓練期間は，2013/10/2 から各テスト期間前までを設定する．

3.3.2 評価指標

インフルエンザ流行予測モデルの評価指標として，決定係数 R^2，平均絶対誤差 MAE，平均絶対パーセント誤差 $MAPE$ の 3 つを用いる．R^2 は回帰方程式の当てはまりの良さを示す指標であり，最大値 1 をとり，値が大きければ良いモデルであることを示す．一般的に，0.5 を超えていることが有用なモデルの目安とされる．MAE は予測値と正解値との絶対誤差の平均をとった指標であり，値が小さければ誤差が少ないモデルであるという指標となる．$MAPE$ は正解値と比較して予測値にどれほどのパーセンテージ誤差が存在するかを表す指標となる．値が小さければより安定したモデルであることを示す．これらの評価指標の計算式を以下に示す．

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (F_t - A_t)^2}{\sum_{t=1}^{n} \left(A_t - \overline{A_t}\right)}$$

$$MAE = \frac{\sum_{t=1}^{n} |F_t - A_t|}{n}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{F_t - A_t}{A_t} \right| * 100(\%)$$

このとき，n はサンプル数，A_t を時点 t における正解値，F_t を時点 t における予測値を示す．インフルエンザ罹患者数の予測においては流行期間の予測が重要であり，モデルの適合具合を評価する決定係数 R^2 が非流行期よりも流行期の影響を強く受けることから，本論文では R^2 が最も重要な評価指標と考える．

3.3.3 結果

各モデル・各期間でインフルエンザ罹患者数予測を行った実験結果を表 3.2 に示す．この結果が示すように，Huber 回帰を用いたモデルが R^2 と MAE において最も高い予測精度を達成している．一方で，$MAPE$ においては Random Forest を用いたモデルが高い精度となった．図 3.3 と図 3.4 に 4 つのモデルを用いたインフルエンザ罹患者数予測の時系列を示す．

表 3.2　インフルエンザ流行予測モデルの精度比較

		Lasso 回帰	Huber 回帰	SVR	Random Forest
2015/05	R^2	0.488	**0.907**	0.693	0.718
–	MAE	30418.20	**8457.69**	15083.25	13857.41
2016/05	$MAPE$	3004.88	242.54	199.74	**105.23**
2016/05	R^2	0.676	**0.889**	0.887	0.864
–	MAE	19138.73	**6359.84**	12113.98	10686.23
2017/05	$MAPE$	1801.58	85.93	**76.17**	117.93
2017/05	R^2	0.572	**0.917**	0.619	0.823
–	MAE	24251.58	**7949.70**	22192.46	12240.59
2018/05	$MAPE$	433.45	42.94	116.68	**39.19**

評価指標の性質から考えると，R^2 が高い精度であることから Huber 回帰を用いたモデルが流行期間（12 月-3 月）においては他のモデルよりも罹患者数予測モデルとして適しているといえる．一方で，Random Forest を用いたモデルの $MAPE$ が高いことから，流行期間や非流行期間といった期間に関わらず，安定した予測が可能であるといえる．

3.4　考察

3.4.1　各モデル

通常の線形回帰から損失関数を変更した Huber 回帰が 2 つの評価指標 R^2 と MAE において最も高い精度となった．Huber 回帰は目的変数の外れ値に対し頑健であるという性質を持ち，同様に説明変数の外れ値に対しても一定程度の頑健性を持つことが知られている．今回の実験で利用した多くのリソースの中でも，Twitter におけるクローリングの不調による取得ツイート数の減少や検索クエリのアクセス過多による急激な検索数の上昇といった問題によって観測値の急激な変化が見られた．上記のような問題によって，他のモデルと比較し Huber 回帰によるモデルが良い結果をもたらしたと考えられる．Web 上のリソースを用いる場合，アクセス数の急激な増加やクローリング手法の変更・停止などの問題が常に存在する．このこ

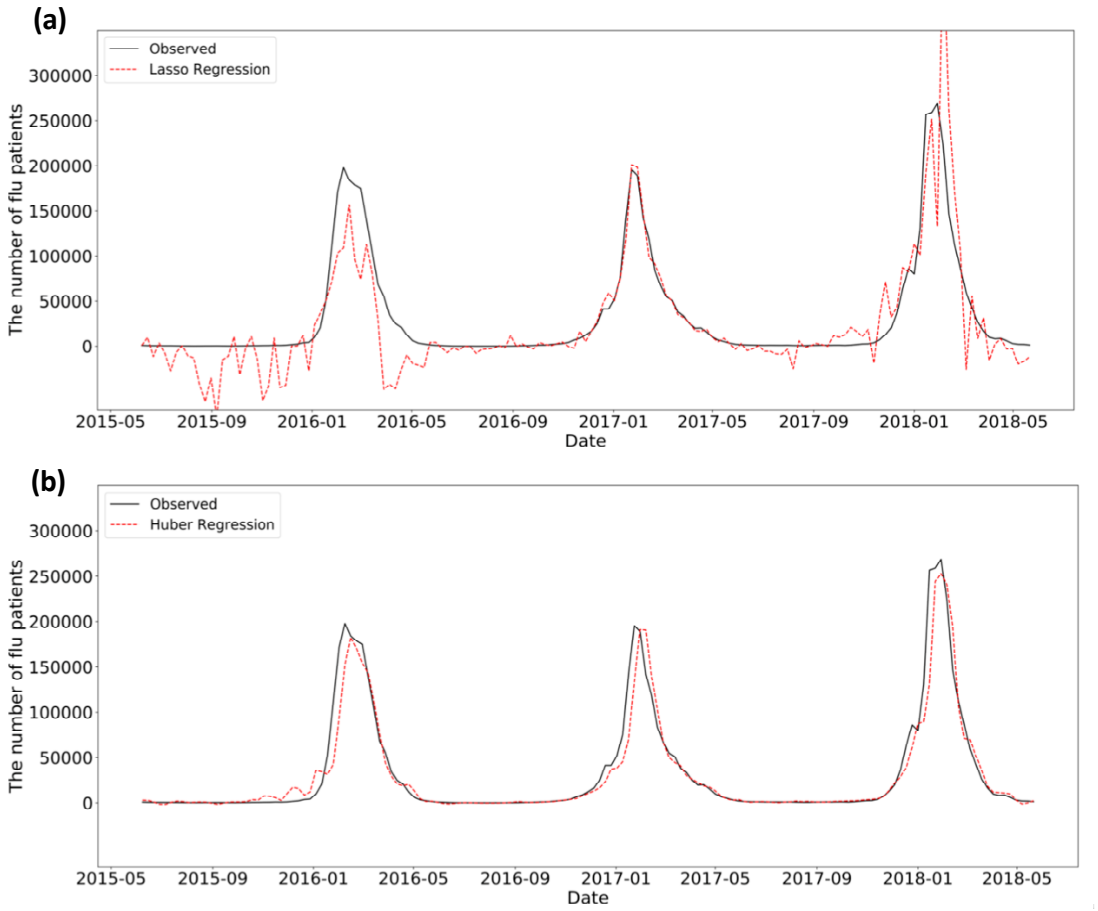


図 3.3　(a) Lasso 回帰，(b) Huber 回帰によるインフルエンザ流行予測の時系列

とからも，インフルエンザなどの感染症流行を予測する頑健なモデル作成に Huber 回帰を用いることは効果的であると考えられる．

一方で，Random Forest を用いた流行予測モデルは，$MAPE$ において Huber 回帰よりも高い精度を達成している．$MAPE$ は正解値が小さいときに予測結果で多くズレると評価指標の結果が極端に悪化するという特徴がある．このことからも，日本などの温帯地域で非流行期・流行期の流行の大きさが極端に異なる場合，非流行期において安定した出力がされることが直接 $MAPE$ の精度に繋がる．また，Random Forest を用いたモデルと比較して，入力値が大きく変動する可能性がある本実験のリソースと回帰によるモデルでは，非流行期に安定した予測の出力がされないという傾向がある．よって，Random Forest によるモデルは流行期・非流

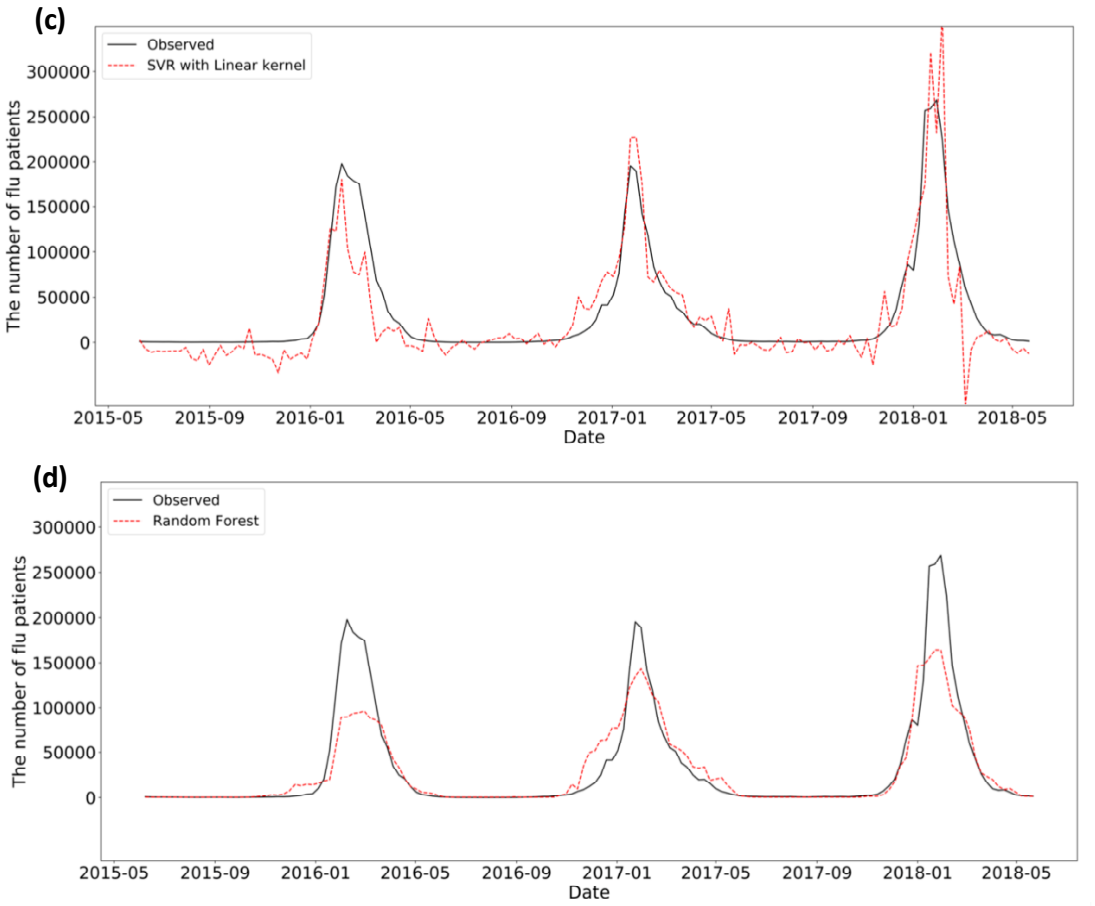


図 3.4　(c) SVR，(d) Random Forest によるインフルエンザ流行予測の時系列

行期に関係なく安定した予測値を出力し，$MAPE$ では良い結果になったと考えられる．

インフルエンザ罹患者数予測モデル作成において，非流行期よりも流行期の予測の当てはまり具合の良さの方が重要であることから，Huber 回帰を用いたモデルが実用上はより適していると考えられる．

3.4.2　各リソース

Huber 回帰を用いたモデルを使用して，本実験で用いたリソースから 1 つずつ除いて同様の実験を行い，それぞれのリソースの予測性能の貢献度の検証を行なっ

た．その結果を表 3.3 に示す．2017/05–2018/05 においては Twitter を除いたモデルが最も高い精度を達成するが，これは期間中にツイートクローリングが上手く行かなった時期があり，それが要因であると考えられる．また，過去の罹患者数を特徴から抜いたモデルでは大幅に精度が下がることから，重要な特徴であることが示唆される．一方で，過去の罹患者数以外の特徴においては，年度ごとに異なる有用性を示しており，一貫としたインフルエンザ流行予測モデル作成の難しさを示している．

過去の罹患者数が特に有用な特徴であるということは，図 3.5 からも確認される．図 3.5 は，実際のインフルエンザ罹患者数の値（黒），Huber 回帰による予測値（赤），そして，過去の罹患者数のみを特徴とする Autoregressive integrated moving average (ARIMA) Model を用いた際の予測値（青）の時系列を表す．表 3.4 に $ARIMA(3,1,2)$ を用いた精度を示す．$ARIMA(p,d,q)$ は，

$$\left(\sum 1 - \sum_{i=1}^{p} \phi^i L^i\right)(1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\epsilon_t$$

というモデルである．ここで，ϕ_i は AR 部分のパラメータ，θ_i は MA 部分のパラメータ，ϵ_t はエラー項，L はラグオペレーターを示す．

図 3.5 と表 3.4 からわかるように，過去の罹患者数のみを用いたモデルでも十分な精度が達成された．当然ながら，単一リソースを用いたモデルより全てのリソースを含んだモデルが高い精度を達成するが，わずかに精度が向上するモデルのために全てのリソースを用意するためのコストをかける必要があるかどうかは実用上議論の余地があるだろう．

これらの結果から，発展途上国など公衆衛生機関や情報が不十分な国にとっては，ソーシャルメディアの投稿ログや検索クエリなどの UGC が特に有用となるといえる．一方で，公衆衛生情報が十分な国にとって，インフルエンザのような季節性の強い感染症の流行予測のために UGC が有用となる機会は限られるものの，予測タイムラグの修正や突然の流行検知などの特定の機会には有用である．

表 3.3　各リソースを除いた Huber 回帰を用いた予測精度比較

		all	w/o 検索クエリ	w/o SNS 投稿	w/o ショッピング検索	w/o Q&A サービス	w/o 過去の罹患者数
2015/05	R^2	**0.907**	0.906	**0.907**	0.906	0.906	0.489
–	MAE	**8457.69**	8637.29	8573.25	8603.04	8569.50	19065.75
2016/05	$MAPE$	**242.54**	254.78	263.66	255.19	249.81	302.54
2016/05	R^2	0.889	0.903	0.880	**0.904**	0.881	0.448
–	MAE	**6359.84**	6419.06	6754.79	6887.97	6443.73	18917.90
2017/05	$MAPE$	85.93	110.41	**59.62**	110.09	73.03	710.61
2017/05	R^2	0.917	0.900	**0.921**	0.902	0.910	0.449
–	MAE	7949.70	8493.72	**7303.47**	8258.65	8024.62	25941.20
2018/05	$MAPE$	42.94	45.59	**32.32**	37.09	37.75	173.73

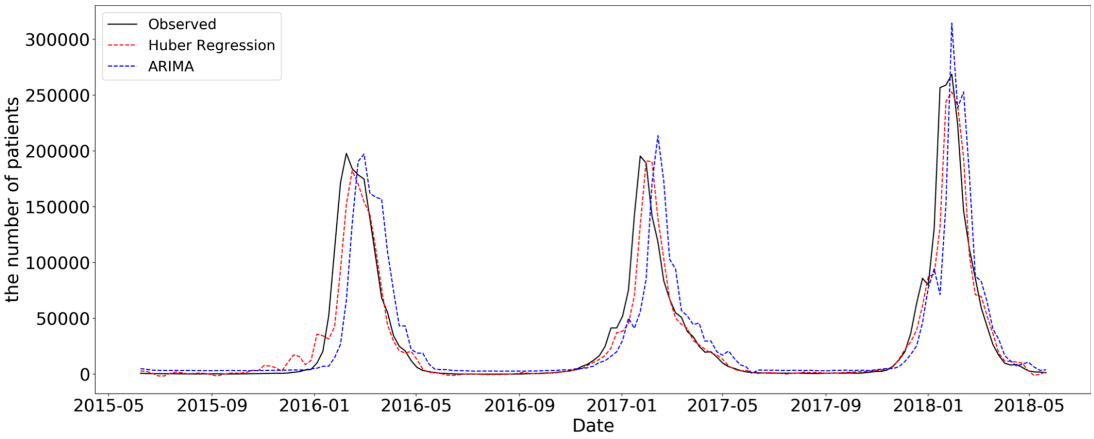


図 3.5　2015/05–2018/05 でのインフルエンザ罹患者数，Huber 回帰による予測，ARIMA モデルによる予測の時系列

3.4.3　変化点分析

インフルエンザ罹患者数の時系列の動きに対して，各リソースの時系列の動きが同様に変動もしくは，前もって変動していると予測のためのリソースとして有用であると考えられる．このことから，Barry and Hartigan's Bayesian 変化点分析の手法 [39] を用いて時系列の変化が同様のタイミングで起こっているかどうか分析する．

Barry and Hartigan's Bayesian 変化点分析は時系列中の大きな変化を検出するために，金利データ [41] や癌関連遺伝子発見 [40]，インフルエンザ検知の観点からも Wikipedia や Twitter などのリソースを用いたインフルエンザ検知 [7] など，

表 3.4　ARIMA を用いたインフルエンザ罹患者数予測精度

		ARIMA
2015/05	R^2	0.823
–	MAE	12953.98
2016/05	$MAPE$	57.77
2016/05	R^2	0.723
–	MAE	11472.42
2017/05	$MAPE$	131.01
2017/05	R^2	0.733
–	MAE	16250.42
2018/05	$MAPE$	61.69

多くの分野で用いられている手法の 1 つである．本手法では，観測値は全て独立 $N(\mu_i, \sigma^2)$ であると仮定し，各点 i における変化点確率 $p \in [0,1]$ が求められる．時系列における区切りを $\rho = (U_1, U_2, ..., U_n)$ とする時，$U_i = 1$ は時点 $i+1$ において変化していることを示す．

時系列における時点 i での変化点確率 p_i は以下の式で求まる．

$$\begin{aligned}\frac{p_i}{1-p_i} &= \frac{P\left(U_i = 1|\mathbf{X}, U_j, j \neq i\right)}{P\left(U_i = 0|\mathbf{X}, U_j, j \neq i\right)} \\ &= \frac{\left[\int_0^\gamma p^b \left(1-p\right)^{n-b-1} dp\right]\left[\int_0^\lambda \frac{w^{b/2}}{(W_1+B_1 w)^{(n-1)/2}} dw\right]}{\left[\int_0^\gamma p^{b-1} \left(1-p\right)^{n-b} dp\right]\left[\int_0^\lambda \frac{w^{(b-1)/2}}{(W_0+B_0 w)^{(n-1)/2}} dw\right]}\end{aligned}$$

このとき，W_0，B_0，W_1，B_1 はそれぞれ $U_i = 0$，$U_i = 1$ のときの区間同士の二乗和 (B) と区間内の二乗和 (W) を示す．λ，γ はハイパーパラメータで $[0,1]$ の値を取る．j を最後の点，b を $P\left(U_i = 0|U_j, j \neq i\right)$ のときの区間数を示す．MCMC 近似を用いて W_0，B_0，W_1，B_1 が更新され，積分が計算される．

インフルエンザ罹患者数の時系列の変化点とその他の UGC リソースの変化点が一致するかどうかを，上記の手法を用いて確認する．遷移確率 p が 50% を超えるものを変化点とみなす．また，UGC がインフルエンザ罹患者数の変化に対して，早期にもしくは遅く検知している可能性があることから，インフルエンザ患者の時系列の変化点の 1 週間前から 1 週間後までに UGC の変化点が発生した場合に一致しているとみなし，評価指標 $Sensitivity$ と PPV を測定する．インフルエンザ

罹患者数と UGC リソースで一致したとみなされた変化点を真陽性 (true positive) とみなす．インフルエンザ罹患者数の時系列では変化点が検知されている一方で，UGC リソースでは検知されていない変化点を偽陰性 (false negative) とする．また，インフルエンザ罹患者数の時系列では変化点が検知されていない一方で UGC リソースでは検知されている変化点を偽陽性 (false positive) とする．このとき，$Sensitivity$ と PPV は以下のように定義される．

$$Sensitivity = \frac{true\,positive}{true\,positive + false\,negative}(\%)$$
$$PPV = \frac{true\,positive}{true\,positive + false\,positive}(\%)$$

MCMC の iteration 回数を 500，ハイパーパラメータ λ，γ をそれぞれ 0.1 と設定し，各リソースにおいてインフルエンザ罹患者数との相関が高い上位 3 クエリを用いて変化点検知を行なった．

結果を表 3.5 に示す．2016/05–2017/05 に限り，取得したツイート量が不安定だったことから 2 つの指標において不安定な値となっている．しかし，それを除いて検索クエリと SNS 投稿は双方ともいずれの評価指標においても高い値を保っており，インフルエンザ罹患者数の予測モデルに対して安定して貢献する予測因子であるといえる．

ショッピング検索では PPV は高く，$Sensitivity$ は低いという結果となった．これは，ショッピング検索における検索数の変化点がインフルエンザ罹患者数の変化点と一致する一方で，全ては捉えきれていないということを示している．

一方で，Q&A サービスの検索では $Sensitivity$ が高いが PPV が低いという結果となった．これは，Q&A サービスの検索数はインフルエンザ罹患者数の変化点を捉えているが，ノイズが多いため，インフルエンザ罹患者数の時系列の変化点でない部分でさえ変化点として捉えられたことが伺える．

2014/11/10–2015/03/22 の各リソースの変化点の時系列を図 3.6 に示す．各リソースは [0,1] に正規化され，各リソースの時系列と同じ色の縦線は，そのリソースで変化が検出された変化点を示す．

表 3.5　変化点検知におけるインフルエンザ罹患者数の時系列と UGC リソースの変化点合致率

		検索クエリ	SNS 投稿	ショッピング検索	Q&A サービス
2015/05	$PPV(\%)$	60	60	38	55
–2016/05	$Sensitivity(\%)$	43	76	56	32
2016/05	$PPV(\%)$	73	8	25	31
–2017/05	$Sensitivity(\%)$	57	33	60	25
2017/05	$PPV(\%)$	80	75	47	47
–2018/05	$Sensitivity(\%)$	75	71	78	57

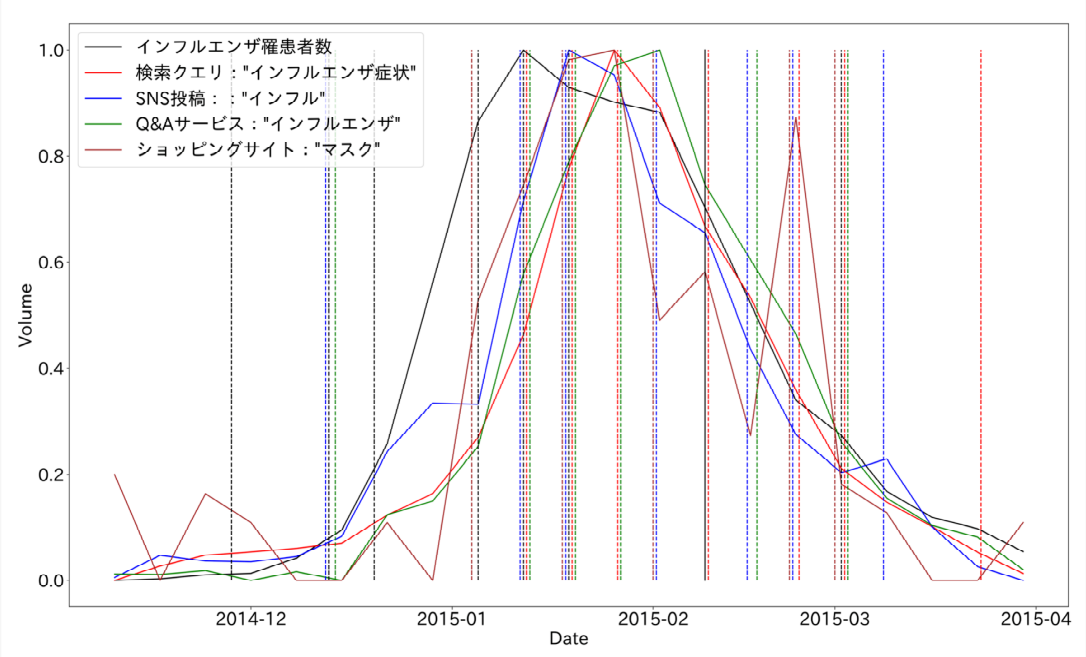


図 3.6　2014/11/10–2015/3/22 における変化点の時系列．各リソースは最小 0，最大 1 になるよう正規化されている．各リソースの時系列と同じ色の縦線は，そのリソースで変化が検出された変化点を示す．

3.5　本章のまとめ

本章では，5 つのリソース：過去の罹患者数，検索クエリ，SNS 投稿，ショッピング検索，Q&A サービスを用いてインフルエンザ流行予測モデルを作成し，各リソースの有用性について検討を行った．本章の実験を通して以下の洞察が得られた．

- Huber 回帰は UGC を用いたインフルエンザ流行予測に対して頑健であり，精度としても高い結果をもたらす．
- 複数のリソースを用いたアンサンブルモデルは，高い予測結果をもたらす．
- 過去の罹患者数は重要な特徴量である．
- インフルエンザの流行において，ショッピング検索量にはほとんど変動が反映されず，Q&A サービス検索量はノイズが多い．

本研究の結果は，インフルエンザ流行予測のために検索クエリと SNS 投稿の有用性は年に依存し，大きな違いを示さないことが示唆された．確かに，複数のリソースを用いたアンサンブルモデルは信頼性が高く，頑健な予測を行うが，全ての複数リソースを用いることによる精度向上は小さい．これら全てのリソースを用意するコストは高いことからも，十分な公衆衛生情報を持つ国では，過去のインフルエンザ罹患者データに加えて単一の UGC（検索クエリまたは Twitter）を使用することがコストも安く，正確な予測モデルの作成が可能だと考えられる．

次章では単一の UGC (Google Trends) と過去のインフルエンザ罹患者データを用いて新たな予測モデルの提案を行う．

第 4 章　モデル：レギュラー部分とイレギュラー部分を結合したインフルエンザ流行予測

4.1　本章の概要

これまで，インフルエンザ流行予測に対して Online user-generated contents (UGC) とインフルエンザ罹患者データの 2 つのリソースが主に用いられてきた．前章でも述べたように，UGC においては SNS 投稿 (Twitter) [11] や検索クエリログ (Google) [3] などが主な特徴量として用いられ，流行予測が行われている．UGC を予測の特徴量として用いることの利点として，突然の流行発生をすばやく検知できることが挙げられる．しかしながら，UGC を用いたインフルエンザ流行予測においては，以下の問題が存在する．

1. 過去のパターンなど固有の情報を捉えることの難しい [10]
2. UGC におけるノイズによって，非流行期での予測値が大きく変動する可能性がある

一方で，インフルエンザ罹患者データは流行の季節性などの情報が明示的に含まれており，前章で述べたようにインフルエンザ罹患者データのみをリソースとして用いたモデルでも十分に流行予測が可能である．しかしながら，インフルエンザ罹患者データのみをリソースとしたモデルでは過去のパターンを学習するのみで，例年と異なる突発的な流行や変動を捉えることは難しい．より有用なインフルエンザ流行予測モデルにおいては，

1. 季節性など時系列固有の部分 (regular trend)
2. 突然の予期できない変動部分 (irregular trend)

の 2 つの異なる現象を把握・予測する部分に分解できる．リソースとしてインフルエンザ罹患者データは regular trend を把握し，UGC は irregular trend を把握することに利用できる．上記の利点を把握し，これらのリソースを同時に用いて予測モデルの作成を行った研究 [8, 9] は存在するが，それぞれのデータを区別し利点を的確に考慮したモデルは未だ考案されていない．

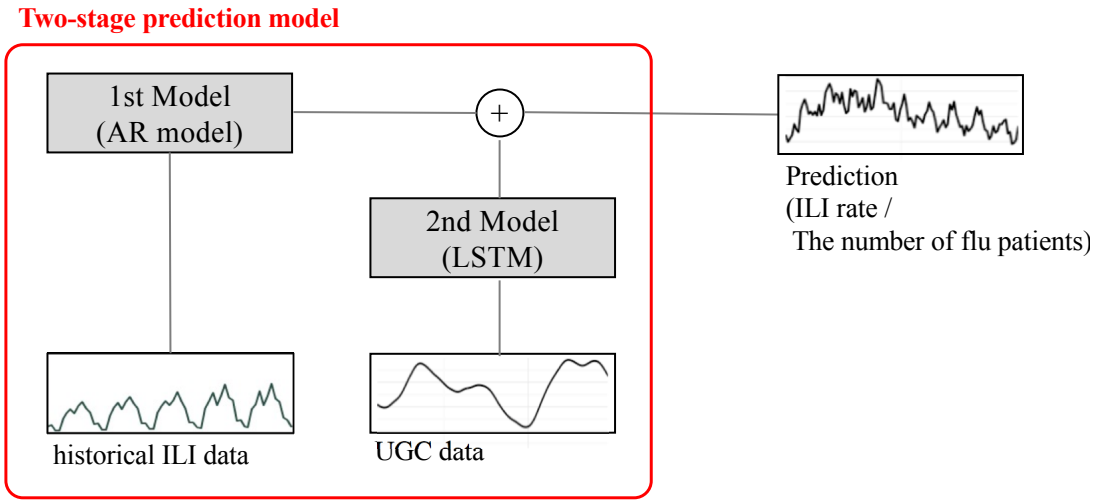


図 4.1　Two-stage Model の概要：過去のインフルエンザ罹患者データと UGC データの 2 つを個別のモデルに入力して学習を行い，予測の際に 2 つの出力を組み合わせる．Two-stage Model は通常のトレンドを予測する 1st Model とイレギュラーなトレンドを予測する 2nd Model の 2 つのモデルから構成される．

本章では 2 つのデータ（インフルエンザ罹患者データと UGC）の特性を踏まえて，regular 部分と irregular 部分を明示的に組み合わせる手法を提案する．2 つのデータを単に同時に用いるのではなく，インフルエンザ罹患者データを用いたモデルと UGC を用いたモデルを別々に作成する．インフルエンザ罹患者データを用いたモデル (1st Model) は時系列の規則的なパターン (regular trend) を予測し，UGC データを用いたモデル (2nd Model) は突然の変動 (irregular trend) 部分を予測する．この 1st Model と 2nd Model を組み合わせたモデルを Two-stage Model とする．Two-stage Model の概要を図 4.1 に示す．

4.2　提案モデル: Two-stage model

4.2.1　モデル概要

利点の異なる 2 つのデータ，過去のインフルエンザ罹患者データと UGC データを用いて，インフルエンザ流行予測モデルを作成する．各データの有用な点を活用するために，予測のためのプロセスを 1st Model と 2nd Model の 2 つに分割する．

本章では 1st Model と 2nd Model を組み合わせたモデルを Two-stage Model として提案する．

1st Model

1 つ目のプロセスでは，Auto Regression (AR) モデルを用い，過去のインフルエンザ罹患者データから通常の変動を学習し，予測モデルを作成する．本章では，UGC データを用いた 2 つのモデルのベースとなることから，このプロセスで用いるモデルを 1st Model と呼ぶ．

2nd Model

2 つ目のプロセスは突然の変動や流行を学習・予測するためのものである．インフルエンザ罹患者データから学習したモデルでは予期できない変動を学習するために，実際の罹患者数（正解値）と 1st Model における予測値の差分を 2 つ目のプロセスにおける正解値として学習を行なう．つまり，1st Model が予測を外した部分の学習を 2 つ目のプロセスで行なう．

UGC データの変動が同様に突然の変動や流行を表していると仮定し，特徴量として UGC データの検索量をそのまま扱うのではなく，前の時点からどれだけの変動が生じたかを示す差分量を用いる．本章では UGC データを用いたプロセスを 2nd Model と呼ぶ．

図 4.2 に 1st Model と 2nd Model の概要を示す．

4.2.2 実装

本章では，モデルの簡易化と比較のために 1st Model に AR モデル，2nd Model に Long-Shot Term Memory (LSTM) を用いる．

AR モデル

AR モデルは確率過程の 1 種で，時系列分析に用いられる手法である．自己相関の高い時系列データに用いられ，前の時点の値を用いて予測値を算出される．

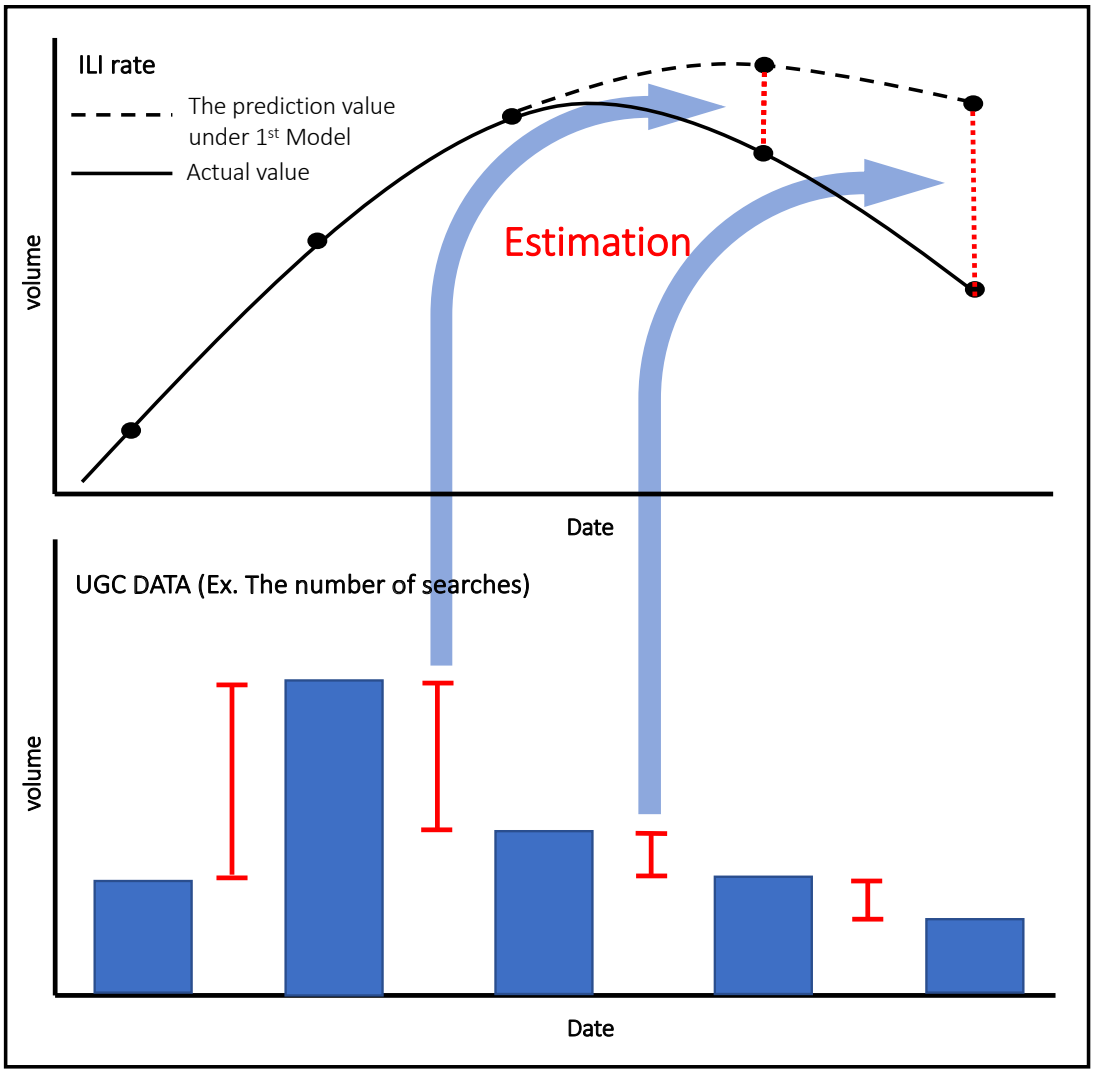


図 4.2　1st Model は過去のインフル罹患者データから学習を行い，2nd Model で 1st Model の予測値と実際の罹患者数との差分を UGC を用いて予測を行なう

$AR(o)$ は以下の式で表される．

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t$$

ここで p は現在の値を予測する際に考慮する過去の時系列の長さ， φ_i はモデルにおける説明変数，c は定数，ε_t はホワイトノイズを表す．

Long Short-Term Memory (LSTM)

LSTM は RNN (Reccurent Neural Network) の一種であり，多くの NLP や音声処理などの問題に適用されているモデル [42, 43] の 1 つである．RNN は時系列における依存関係を捉えられることが特徴であり，時点 i における正解値 y_i を予測するために，前の入力 x_1, ..., x_i を考慮する．しかしながら，入力する時系列データの長さが長い際，伝搬する誤差が過剰に大きく，もしくは小さくなるため，長期的な依存関係を学習することが難しいという問題がある．LSTM はセル構造を持つことで，深い層においても勾配を計算し長期的な依存関係を学習することができる．LSTM の構造は現在の入力を使用するかを選択する入力ゲート，過去のセル情報を忘れるかを決定する忘却ゲート，どれだけの情報を出力するかを選択する出力ゲートの 3 つからなる．シグモイド関数を用いることで [0,1] の範囲で値を取り，それぞれのゲートがセルに値を保持するか，廃棄するかの選択を行なう．それぞれのゲートと隠れ層は以下のように求められる．

$$
\begin{aligned}
i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\
o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \sigma\left(W_c x_t + U_c h_{t-1} + b_c\right) \\
h_t &= o_t \odot \sigma\left(c_t\right)
\end{aligned}
$$

i_t，f_t，o_t はそれぞれ入力ゲート，忘却ゲート，出力ゲートとする．x_t は LSTM の入力，h_t は隠れ層，c_t はメモリセル内の状態，$\odot$ は要素ごとの積を表す．

4.3 実験

日本とアメリカの 2 カ国を対象に，Two-stage Model の予測性能を比較するための実験を行なった．それぞれの国の，予測リソースと正解値として用いるインフルエンザ罹患者データと UGC データを収集する．UGC データは検索量の取得が可能な Google Trends*を用いる．

*https://trends.google.com

4.3.1 アメリカのデータセット

インフルエンザ罹患者データ

米国の公的感染症対策組織である CDC (Centers for Disease Control and Prevention) はインフルエンザ様疾患 (Influenza like illness，ILI) に関するレポートを毎週公開している．日本の定点辺りのインフルエンザ罹患者数と異なり，アメリカではインフルエンザと考えられる症状を ILI として集計している．CDC による報告は集計のため約 1 週間遅れで公開され，Fluview[†]から無料で取得可能である．アメリカのインフルエンザ流行予測研究では，地域ごとの医療機関に対する ILI 罹患者数の規模をパーセントで表した weighted ILI rate が予測の対象として用いられることが多い．このことから，本章の実験でも 2010/07/10 – 2018/09/08 における毎週の weighted ILI rate をアメリカのインフルエンザ罹患者データとして用いる．

Google Trends

2nd Model におけるリソースとして Google Trends のデータを用いる．具体的には，先行研究 [8] で用いられたクエリの中から 20 クエリを特徴としてランダムに選択し，2012/09/01 – 2018/09/08 における各特徴の Google での検索量を取得する．

4.3.2 日本のデータセット

インフルエンザ罹患者データ

3 章と同様に，NIID の提供しているデータを用いる．2011/09/11 – 2018/04/09 までの定点あたりのインフルエンザ罹患者数を利用する．

Google Trends

3 章の検索エンジンのクエリ選択手法と同様に，2013/10/02 – 2018/05/21 の「インフル」もしくは「インフルエンザ」を含んだツイートのうち，高頻出語上位

[†]http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

表 4.1　本実験で設定した訓練期間とテスト期間

期間	アメリカ	日本
1st Model 訓練	2010/07/10 – 2012/08/31	2011/09/11 – 2013/10/06
2nd Model 訓練	2012/09/01 – 2017/06/03	2013/10/07 – 2017/02/05
テスト	2017/06/04 – 2018/09/08	2017/02/06 – 2018/04/09

100 語と tf-idf の値が上位 50 語を候補語として抽出する．訓練期間中のインフルエンザ罹患者データと候補語の Google Trends の検索量の相関上位 20 語を特徴として用いるクエリとして選択し，2013/10/07 – 2018/04/09 までの検索量を用いる．

4.3.3　設定

Two-stage Model の有用性をアメリカと日本の 2 カ国におけるインフルエンザ罹患者数の予測を行い評価する．ここでの予測とは 1.2 節で述べた予測と同様に，i 番目の週 t_i を現時点とすると，その一週間後 t_{i+1} のインフルエンザ罹患者数もしくは ILI rate の予測を行なうことである．1st Model の入力として 3 週間分のインフルエンザ罹患者データを用いる．つまり，1st Model は AR(3) として学習を行なう．2nd Model では，合計 2 週間から 52 週間（約 1 年間）の Google Trends データを入力としてモデルの作成を行ない，隠れ層を 80，層数を 1 として学習を行なう．Two-stage Model に対してのリソースの用い方を図 4.3 に示す．

アメリカのインフルエンザ流行予測モデル作成において，1st Model の訓練期間を 2010/07/10 – 2012/08/31, 2nd Model の訓練期間を 2012/09/01 – 2017/06/03 と設定する．テスト期間は，2017/06/04–2018/09/08 の 66 週間を設定する．日本のインフルエンザ流行予測モデル作成では，1st Model の訓練期間を 2011/09/11 – 2013/10/06，2nd Model の訓練期間を 2013/10/07 – 2017/02/05 と設定する．テスト期間は，2017/02/06 – 2018/04/09 の 62 週間と設定する．表 4.1 に実験期間のまとめを示す．

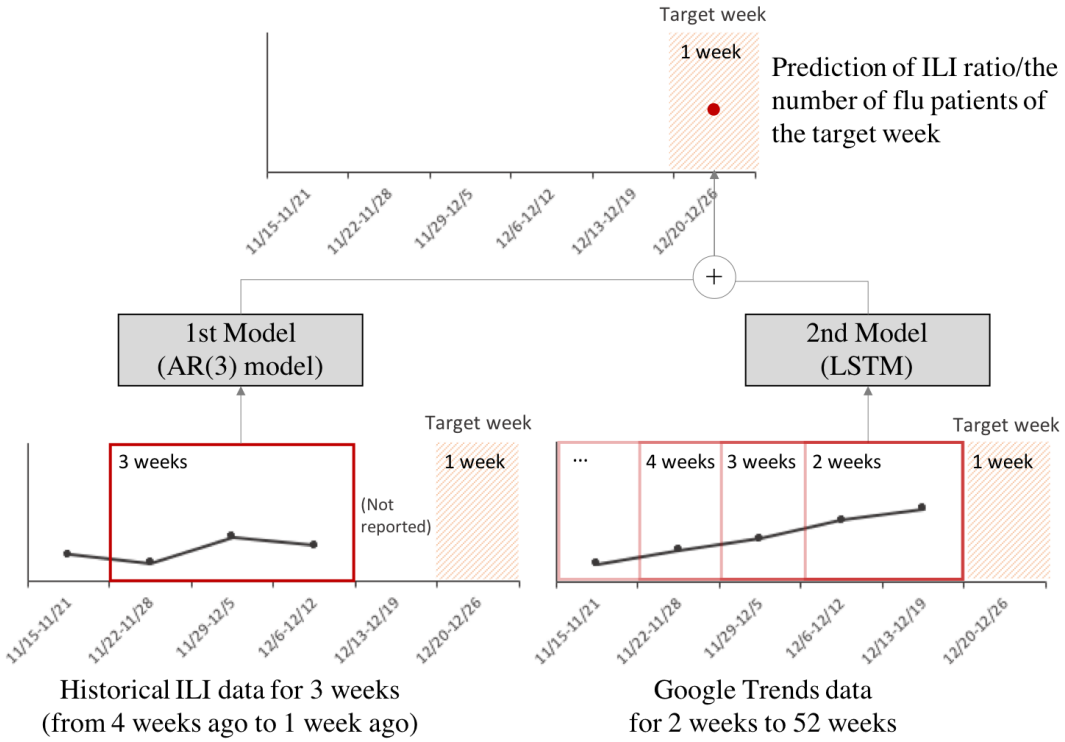


図 4.3　Two-stage Model の予測例：12/20 – 12/26 のインフルエンザ罹患者数を予測する際，11/22（4 週間前）から 12/12 までの過去の罹患者データを 1st Model の入力として用いる．一方で，2nd Model には Google Trends データを最短 2 週間分から最長 52 週間分のデータを入力とし学習を行なう．なお，利用データの長さはモデルに依存する．

4.3.4　比較手法

Two-stage Model を評価するために，本手法と近い考えの元考案された ARGO モデル [8]，様々なリソースを用いたインフルエンザ流行予測に用いられている Random Forest [24, 30]，そして Two-stage Model の最初のプロセスのみの 1st Model Only すなわち AR モデルを比較手法として用いる．

ARGO Model

ARGO モデル [8] は隠れマルコフモデルを元に考案されたモデルで，統計学的には外部リソースとして Google Trend を用いた ARX (Auto Regressive eXogenous

model) モデルである．ARGO モデルは以下の式で定義される．

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \varepsilon_t$$
$$\varepsilon_t \sim N(0, \sigma^2)$$

このとき，y_t は時間 t における ILI rate，$X_{i,t}$ はクエリ i，時間 t の Google Trends における検索量を示す．また，インフルエンザ罹患者データの季節性の動きを捉えるために，N は ARGO モデルの入力として用いる週数を示す．本実験では，$N = 52$(weeks) とする．最適なパラメータ μ_y, $\alpha_1 ... \alpha_{52}$，$\beta_1 ... \beta_{20}$ を見つけるために，

$$\sum_t \left(y_t - \mu_y - \sum_{j=1}^{52} \alpha_j y_{t-j} + \sum_{i=1}^{20} \beta_i X_{i,t} \right)^2 +$$
$$\lambda_\alpha ||\alpha||_1 + \eta_\alpha ||\alpha||_2^2 + \lambda_\beta ||\beta||_1 + \eta_\beta ||\beta||_2^2$$

の最小化を行なう．ARGO モデルは罹患者データと Google Trends のそれぞれのデータから自動的に変数選択を行なうため，L1・L2 正則化を適用する．ここでの，λ_α, λ_β, η_α，η_β は L1・L2 正則化のためのハイパーパラメータを示す．

Random Forest (RF)

Random Forest のモデルは 3 章で用いたものと同様である．詳細は 3.2.2 節に示す．なお，それぞれのモデルのハイパーパラメータは cross-validation によって決定する．

4.3.5 評価指標

インフルエンザ流行予測モデルの評価指標として，R^2，MAE，$MAPE$ を用いる．3 章で用いたものと同様のであるため．詳細は 3.3.2 節を参照されたい．

4.3.6 結果

表 4.2(a) と図 4.4 に各評価指標によるアメリカを対象とした ILI rate の予測精度を示す．同様に，表 4.2(b) と図 4.5 に各評価指標による日本を対象としたインフルエンザ罹患者数の予測精度を示す．各表には Two-stage Model の 2nd Model の入力週が 10 週，30 週，50 週のときの結果，各評価指標において最高精度のとなった週 (アメリカでは 51 週，日本は 36，37，44 週) の結果，ならびに Two-Stage Model の平均スコアを載せる．

表 4.2(a) が示すように，アメリカでは Two-stage Model が R^2 と MAE で最も高い精度を達成した．しかしながら，$MAPE$ においては Random Forest が最も高い結果となった．

一方，表 4.2(b) によると，日本では Two-stage Model が各評価指標において最も高い精度となっており，平均スコアも同様に他のモデルよりも高い精度となった．

4.4 考察

4.4.1 各モデル

Two-stage Model が日本とアメリカ両国において，R^2 と MAE の 2 つの評価指標で最も高い精度を達成した．特に日本においては，Two-stage Model の平均精度でも他のモデルよりも高い結果となっている．さらに，R^2 においては 1st Model Only の結果よりも約 0.1 精度が上昇しており，この結果は 2nd Model の有用性を示している．

ARGO モデルは，アメリカにおいて最も高い精度を達成した Google Trends データを 51 週分用いた Two-stage Model に近い精度となっている．一方で，日本における精度はそこまで高くない．この結果は Google Trends などの特徴量の質に大きく依存すると考えられる．

Random Forest は，アメリカで $MAPE$ で最も高い精度を達成しており，同様に日本でも高い結果となっている．これは 3 章での Random Forest の結果が示すように，他モデルと比較して非流行期間における予測値が安定していることが要因だと考えられる．しかしながら，予測が重要な流行期間における誤差は他のモデル

表 4.2　Two-stage Model，ARGO モデル，Random Forest，1st Model Only の精度比較

(a) アメリカ

models	weeks	R^2	MAE	$MAPE$
ARGO	-	0.867	0.457	24.43
RF	-	0.873	0.358	**10.78**
1st Model Only	-	0.784	0.510	21.54
Two-stage Model	10	0.838	0.452	19.11
	30	0.834	0.430	18.52
	50	0.871	**0.349**	15.39
	51*	**0.885***	0.354*	14.68*
	Ave	0.818	0.432	17.85

(b) 日本

models	weeks	R^2	MAE	$MAPE$
ARGO	-	0.698	17267.33	88.17
RF	-	0.824	14539.63	38.39
1st Model Only	-	0.852	10320.12	71.17
Two-stage Model	10*	0.903*	8159.17*	18.61*
	30*	0.912*	9162.59*	19.41*
	36	0.881	8116.62	**17.37**
	37*	**0.948***	7724.66*	19.56*
	44*	0.935*	**7141.88***	18.50*
	50	0.901	8931.13	19.28
	Ave	0.905	8959.52	20.06

アスタリスク (*) は，Random Forest を比較として paired t-test を行った時に統計有意差 ($p < 0.05$) が存在することを示すものである．

と比較しても大きい．そのため Random Forest がインフルエンザ流行予測モデルとして有用となる場合は限定的であると考えられる．

結果として，本章で提案した Two-stage Model が本実験で用いた他モデルよりも，ILI rate もしくはインフルエンザ罹患者数の予測において適していると考えられる．

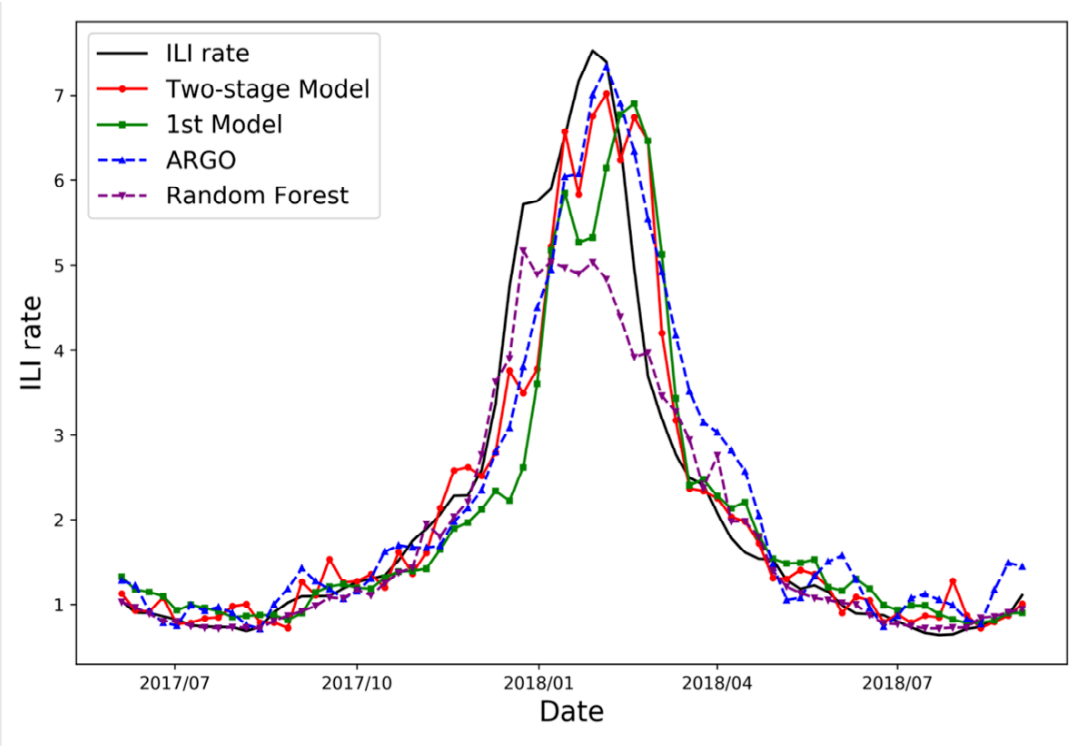


図 4.4　アメリカのインフルエンザ罹患者数予測結果．黒線が実際の値，赤線が提案手法の予測結果を示す．

4.4.2　Two-stage Model の外れ値に対するロバスト性の検証

UGC をリソースとして用いる際，クローリングの不調やアクセス増加などによってデータのノイズや外れ値が多く含まれる例が存在する．これらの問題は予測結果を悪化させる要因の 1 つであり，前処理など時間のかかる工程が要求される．しかしながら，感染症流行予測は緊急性の高さから正確性と迅速さの両方が要求される課題の 1 つであるため，特徴量をそのまま扱えれるように外れ値やノイズに対するモデルのロバスト性が必要となる．Two-stage Model は過去のインフルエンザ罹患者データからの regular trend の予測と，UGC データからの irregular trend の予測という 2 つの部分に分割できる．このことから，一方のデータにノイズが多くとも，もう一方のモデルでノイズの影響を受けずに予測を行なうことができ，Two-stage Model はロバスト性が高いモデルであると考えられる．

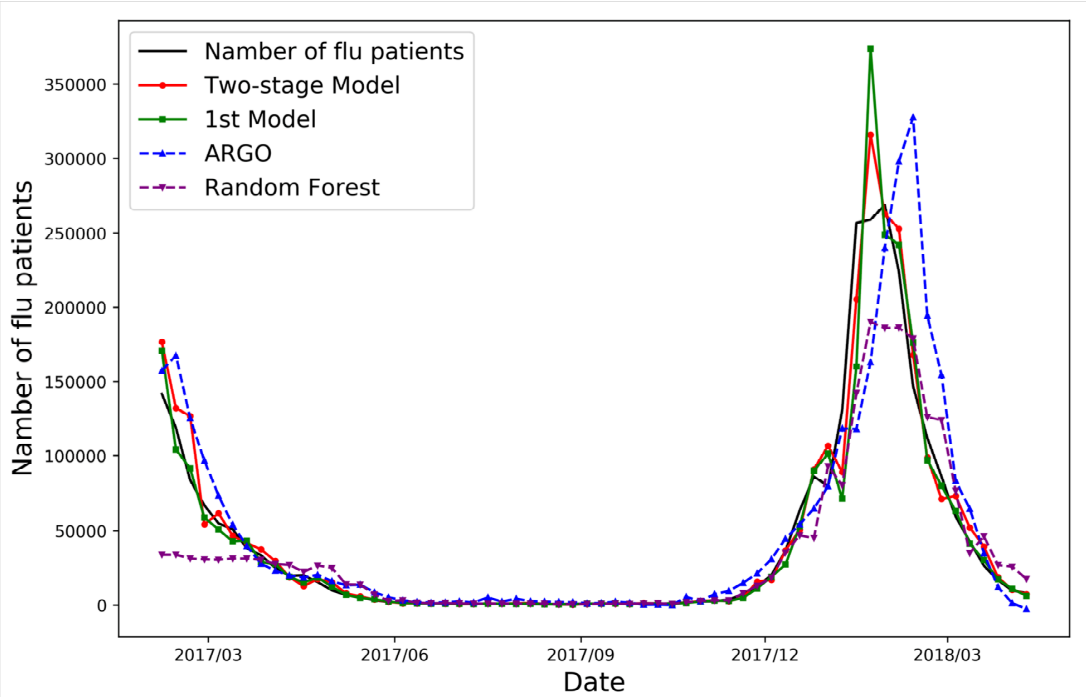


図 4.5　日本のインフルエンザ罹患者数予測結果．黒線が実際の値，赤線が提案手法の予測結果を示す．

モデルが外れ値に対してロバストであるかどうかを検証する．そのために，元の Google Trends データの代わりに意図的にノイズを含めた Google Trends データを用意し，アメリカの ILI rate 予測実験と同じ設定で検証を行なう．元の Google Trends データにノイズを含めるために，最小値 0 から最大値のランダムな値に一部のデータを置き換える．実験のために，元データの 5.0% 10.0%, 15.0%，20.0% をノイズに置換したデータを用いて学習・予測を行なう．

表 4.3 にノイズの割合ごとのモデルの精度結果と，ノイズを含めなかったときとの差分を示す．Two-stage Model の “Best” は 4.3 節のノイズ無しのデータによる実験で R^2 が最も高いスコアだった (表 4.2(a) 参照) Google Trends データを 51 週分用いたモデルのことを指す．

ARGO モデルは，ノイズの割合を増やすにつれて精度が下がっている．一方でノイズを含まない Google Trends データを用いた実験と同様に，Random Forest は $MAPE$ においては他のモデルと比較しても頑健であり，Two-stage Model は

表 4.3　ノイズデータを含めたそれぞれのモデルの精度

Outliers percentage	Metrics	ARGO	RF	Two-stage Model	
				Avg	Best
0%	R^2	0.867	0.873	0.818	0.885
	MAE	0.457	0.358	0.432	0.354
	MAPE	24.43	10.78	17.85	14.68
5%	R^2	0.812 (-0.055)	0.803 (-0.070)	0.808 (**-0.010**)	0.867 (-0.013)
	MAE	0.554 (-0.097)	0.417 (-0.059)	0.469 (**-0.037**)	0.389 (-0.035)
	MAPE	33.25 (-8.82)	14.78 (**-4.00**)	24.72 (-6.87)	18.29 (-3.61)
10%	R^2	0.768 (-0.099)	0.797 (-0.076)	0.799 (**-0.019**)	0.856 (-0.029)
	MAE	0.576 (-0.119)	0.430 (-0.072)	0.497 (**-0.065**)	0.436 (-0.082)
	MAPE	27.04 (**-2.61**)	14.05 (-3.27)	24.29 (-6.44)	22.62 (-7.94)
15%	R^2	0.732 (-0.135)	0.818 (-0.055)	0.801 (**-0.017**)	0.858 (-0.027)
	MAE	0.664 (-0.207)	0.387 (**-0.029**)	0.503 (-0.071)	0.430 (-0.076)
	MAPE	38.75 (-14.32)	12.10 (**-1.32**)	27.59 (-9.74)	25.61 (-10.93)
20%	R^2	0.711 (-0.156)	0.796 (-0.077)	0.793 (**-0.025**)	0.855 (-0.030)
	MAE	0.611 (-0.154)	0.449 (**-0.091**)	0.523 (**-0.091**)	0.425 (-0.071)
	MAPE	28.34 (-3.91)	13.24 (**-2.46**)	21.57 (-3.72)	20.40 (-5.72)

R^2 と MAE において頑健である．Random Forest は 5% のノイズを含めることで大きく精度が下がり，それ以降はノイズの割合を増やしてもほぼ変わらないという傾向がみられる．

一方，データ内のノイズの割合に関わらず，Two-stage Model は流行期における予測の当てはまり具合を示した R^2 において安定しており，下がり幅も-0.010 から-0.025 に留まっている．このことからも，インフルエンザ罹患者数予測を目的とした Two-stage Model は，外れ値やノイズに対して他モデルと比較してロバストであるといえる．

4.4.3　Two-stage Model の強みと弱み

これらの実験と考察を通じて確認された，Two-stage Model の強みは，インフルエンザ罹患者数予測において他モデルと比較して高い精度で可能であること，また UGC で見られる外れ値やノイズに対して頑健なモデルであることである．

Two-stage Model の最も大きな特徴は，予測において regular 部分と irregular 部分の 2 つに分割を行い時系列予測を行った点である．また，1st Model と 2nd Model を時間の流れを解釈できる AR モデルと LSTM モデルを用いた点もこれらの結果を引き起こした要因の 1 つだと考えられる．

今回の実験では比較のために 1st Model に静的なモデル（AR モデル）を用いたが，動的なモデルを用いるとより有用である．例えば，1st Model に動的なモデルの 1 つで時系列における季節性を捉えることができる Seasonal AutoRegressive Integrated Moving Average (SARIMA) [62] Model を用いてアメリカの ILI rate の予測を行った結果を図 4.6 に示す．AR モデルより正解値にフィットしている様子が伺える．R^2 における精度では AR モデルが 0.885，SARIMA モデルを用いたものが 0.950 となった．

さらに，Two-stage Model が 2 つのモデルの組み合わせで構成されていることから，季節性や流行などの傾向が見られる時系列に対してより有用であると考えられる．そのため，本論文では Two-stage Model をインフルエンザ罹患者数予測モデルに対して用いたが，感染症予測に限らず，商品販売予測などのドメインに対してもこの手法を適用できる可能性は高い．

しかしながら，Two-stage Model にも限界や欠点は存在する．まず，Two-stage Model は 1st Model に強く依存しているため，1st Model のみでは予測が出来ない場合などは Two-stage Model でも十分な効果は見られないだろう．そのため，AR モデルのように自己時系列からある程度予測できることが要求される．同様に，本モデルは過去のインフルエンザ罹患者数データなどの公衆衛生データが十分に備わっていることを前提としている．そのため，感染症に関するデータを迅速かつ容易に入手することが困難な国々では，感染症予測のためには Two-stage Model のように過去の罹患者数情報に依存したモデルではなく，前節で述べた複数の UGC を組み合わせたモデルなどの方が予測精度が上がると考えられる．

4.5 本章のまとめ

本章では，1 つの UGC (Google Trends) と過去の罹患者数をリソースとして用いたインフルエンザ流行予測モデルを提案した．アメリカと日本の 2 カ国で，

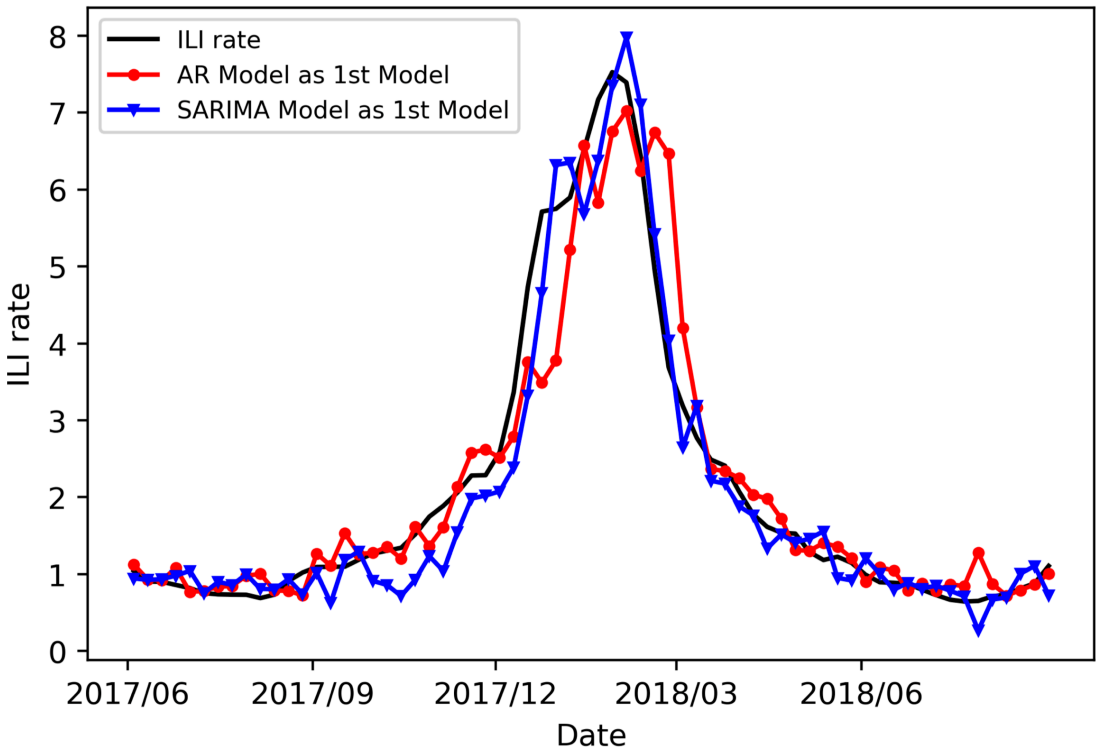


図 4.6　アメリカでの ILI rate 予測：赤は 1st Model が AR(3) モデルのもの，青は 1st Model が SARIMA(3,1,3,52) モデルを用いたものを表す．

Two-stage Model の実験を行い，ベースラインと比較して高い精度でアメリカの ILI 率と日本のインフルエンザ罹患者数の予測ができることを実証した．また，ロバスト検証のための実験を行い，Two-stage Model がベースラインと比較して，外れ値に対してロバストであることを示した．

しかしながら，Two-stage Model は過去のインフルエンザ罹患者データを容易・迅速に入手できることを前提にモデルの構築を行っており，公衆衛生の状況によっては本モデルを適用できない場合も考えられる．このような問題からも，国や場所に応じて適したインフルエンザ流行予測モデルの作成が必要となってくるであろう．本章では Two-stage Model をインフルエンザ流行予測問題に適用したが，商品販売データのように季節性の傾向を示す様々な時系列にも適用可能であると考えられ，本モデルの汎用性は高いと期待される．

第 5 章　エリア：人流を考慮した地域ごとのインフルエンザ流行予測

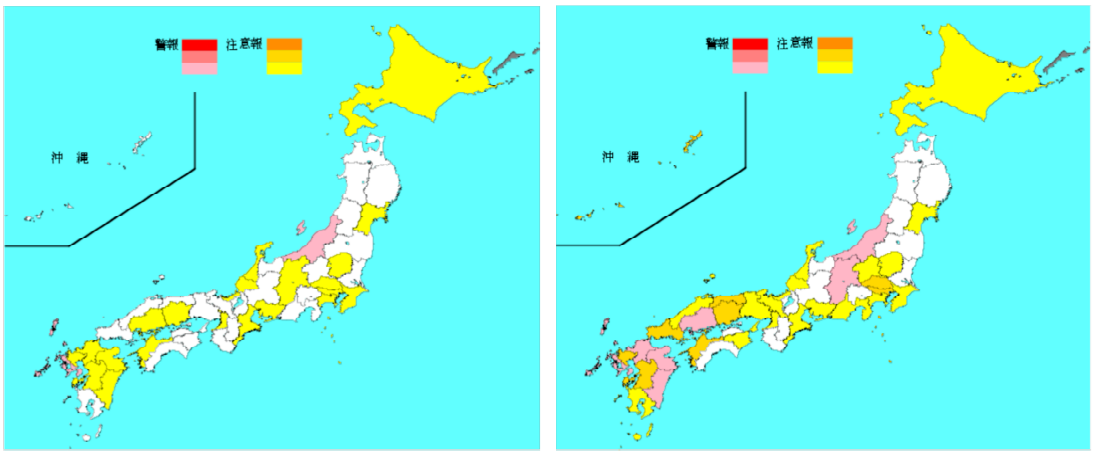
5.1　本章の概要

これまで国単位のインフルエンザ流行予測モデルの作成を行ってきたが，実用上，より詳細な地域ごとの流行予測が求められる．実際に，地域ごとの流行予測の需要は高まっており，最近では米国の公的感染症対策組織である CDC (Centers for Disease Control and Prevention) はエリアごとのインフルエンザ流行予測システム作成を目的としたコンペティションを開催している [64].

インフルエンザなどの感染症の特徴として，宿主の咳やくしゃみによる「飛沫感染」や，宿主の病原体が付着した物体を触ったことによる「接触感染」などにより広がっていき，感染者そのものが流行の原因となることが挙げられる [45]．感染の広がりの例として，一部の感染症は地域から地域へと流行が移っていくという特徴が見られる（図 5.1 参照*）．このことからも，地域ごと（本章では都道府県ごと）のインフルエンザ流行予測を行なう場合，感染が広がっている地域の近くで感染が更に広まっていくという特徴からも，地域ごとの位置関係を考慮し感染症の地域間での広まりをモデルに組み込むことが重要だと考えられる．しかしながら，先程述べた CDC によるコンペティションで提出されたモデルに地域間の位置関係を考慮したモデルは存在せず [46]，多くのモデルは予測精度向上の余地があると考えられる．

これまでの地域間の関係を考慮した感染症予測の研究として，Ransalu ら [47] や Yuexin ら [48] の研究が挙げられ，地域間の関係として位置関係が用いられる．本研究では，感染症が人から人へと感染していく特徴から人流を考慮することが予測においてより有用だと考えられる．このことからも，地域間関係の情報として人流情報がある程度反映された通勤・通学データを用いる．更に，これらの情報を組み込んだ 2 種類のモデル（静的と動的）を用いて，インフルエンザ罹患者予測においてどの程度有効なのかを検証する．最後に，医療機関や公衆衛生機関の決定の助けになるであろ予測区間の付与を行う．

*https://www.niid.go.jp/niid/ja/flu-map.html



(a) 2017年49週　　(b) 2017年50週

図 5.1　インフルエンザの流行地域図：(a) 2017 年 49 週目 (b) 2017 年 50 週目の流行地域を示している．(a) 新潟県，広島県，長崎県などを中心とした流行が，(b) 翌週には周辺の都道府県に広まっている．

5.2　基本モデル

本実験では，地域間の関係を考慮可能な時空間モデルを，入力データが与えられる都度学習を行う動的なモデルと一定期間の学習のみ行う静的なモデルの 2 種類を用いる．動的なモデルとして Spatio–Temporal Neural Network [49] モデル，静的なモデルとして Graph Convolutional Neural Network の一種である Diffusion Convolutional Recurrent Neural Network [50] を用いる．

5.2.1　Spatio–Temporal Neural Network

Spatio–Temporal Neural Network (以下，STNN) モデル [50] は，連続した時空間の依存性を組み込んだ動的なモデルの 1 つとして考案されたモデルの 1 つである．時系列の時空間表現を潜在的に捉える Dynamic 部分と，潜在状態を実際の観測値にデコードする Decoder 部分で構成されている．目的損失関数 $L(d, g, \boldsymbol{Z})$ は以下のように定義される．

$$L(d, g, \boldsymbol{Z}) = \frac{1}{T} \sum_{t} \Delta \left(d \left(\boldsymbol{Z_t} \right), \boldsymbol{X_t} \right) \tag{5.2.1}$$

$$+\lambda\frac{1}{T}\sum_{t=1}^{T-1}||\boldsymbol{Z_{t+1}}-g\left(\boldsymbol{Z_t}\right)||^2 \tag{5.2.2}$$
$$\boldsymbol{Z_t}\in\mathbb{R}^{n\times N},\ \boldsymbol{X_t}\in\mathbb{R}^{n\times M}$$

n を時系列集合の数，M を各系列の次元数，N を潜在表現の次元数としたとき，X_t は時点 t における系列の値，Z_t を時点 t の潜在表現と定義される．$d(.)$ は時点 t における潜在表現 Z_t を実際の値にデコーディングする関数，$g(.)$ は時点 t の潜在表現 Z_t から時点 $t+1$ の潜在表現 Z_{t+1} を予測する関数として定義される．関数 $d(.)$ には MLP を用いるのが一般的である．式 (5.2.1) は Δ は損失関数を表しており，関数 $d(.)$ のデコーディング能力の損失を表現，式 (5.2.2) は関数 $g(Z_t)$ ができるだけ近い潜在表現 Z_{t+1} を学習するための項となっている．目的損失関数を最小化する d，g，Z を探索することによって解が得られる．

$$d^*,\ g^*,\ \boldsymbol{Z^*}=\arg\min_{d,g,\boldsymbol{Z}}L(d,g,\boldsymbol{Z}) \tag{5.2.3}$$

空間情報は関数 $g(.)$ を通してモデルに組み込まれ，$g(Z_t)$ は以下のように定義される．

$$g(\boldsymbol{Z_t})=h(\boldsymbol{Z_t}\boldsymbol{\Theta^{(0)}}+\boldsymbol{W}\boldsymbol{Z_t}\boldsymbol{\Theta^{(1)}}) \tag{5.2.4}$$
$$\boldsymbol{\Theta^{(0)}}\in\mathbb{R}^{N\times N},\boldsymbol{\Theta^{(1)}}\in\mathbb{R}^{N\times N},\ \boldsymbol{W}\in\mathbb{R}^{n\times n}$$

h は非線形関数，$\boldsymbol{W}$ は隣接行列として定義される．$\boldsymbol{\Theta^{(0)}}$ と $\boldsymbol{\Theta^{(1)}}$ は $\boldsymbol{W}$ を用いて依存関係を捉える部分となる．STNN モデルの概要を図 5.2 に示す．

5.2.2 Diffusion Convolutional Recurrenct Neural Network

近年，グラフ構造の学習を行なう Graph Convolutional Neural Network （以下，GCN） [53] が，テキスト分類 [54]，画像解析 [55]，分子構造 [56] などの様々な分野で活用されている．交通量予測問題に対しても GCN を用いた手法や応用法が多く提案されており，自転車の流れ [57]，配車サービスの需要 [58]，道路の交通量 [59] といった予測問題に対して利用されている．

本章の実験では，時間と空間の両方を考慮できるモデルとして，交通量予測のために考案されたモデルの Diffusion Convolutional Recurrenct Neural Network

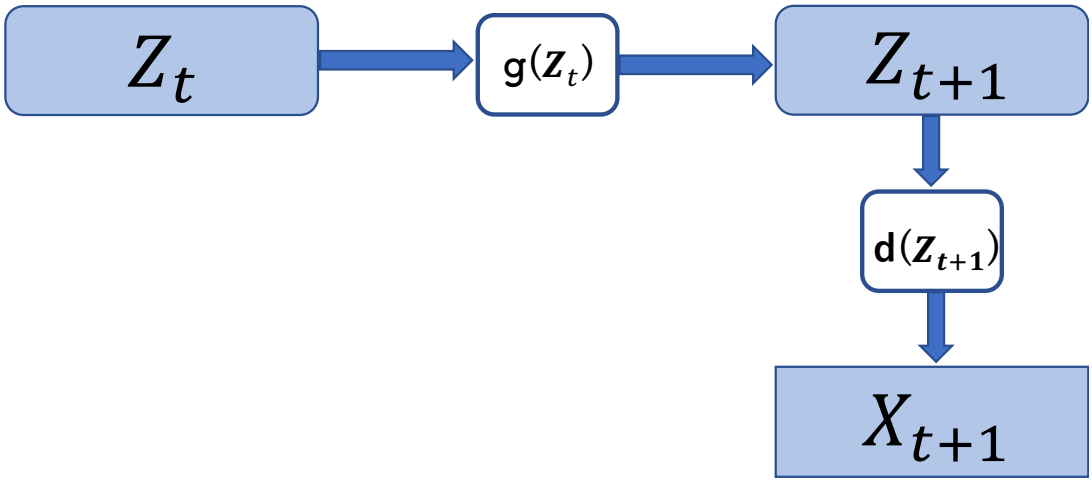


図 5.2　STNN モデルの概要：$d(.)$ を時点 t における潜在表現 Z_t を実際の値にデコーディングする関数，$g(.)$ は時点 t の潜在表現 Z_t から時点 $t+1$ の潜在表現 Z_{t+1} を予測する関数

（以下，DCRNN）モデルを適用する．DCRNN は交通量を拡散過程と関連付けることで空間的依存性をモデル化する．$\boldsymbol{W} \in \mathbb{R}^{n \times n}$ を重み付き隣接行列とし $\boldsymbol{D_O} = diag(\boldsymbol{W1})(\mathbf{1}$ はすべて 1 のベクトル) としたとき拡散過程の定常分布は式 (5.2.5) のように定義される．

$$\boldsymbol{P} = \sum_{k=0}^{\infty} \alpha \, (1-\alpha) \left(\boldsymbol{D_O^{-1}} \boldsymbol{W} \right)^k \tag{5.2.5}$$

式 (5.2.5) を用いて，入力 $\boldsymbol{X} \in \mathbb{R}^{n \times M}$ への Diffusion convolution は

$$\boldsymbol{X}_{:,m} \star \boldsymbol{g} \boldsymbol{f}_{\boldsymbol{\theta}} = \sum_{k=0}^{K-1} \left(\theta_{k,1} \left(\boldsymbol{D_O^{-1}} \boldsymbol{W} \right)^k + \theta_{k,2} \left(\boldsymbol{D_I^{-1}} \boldsymbol{W}^T \right)^k \right) \boldsymbol{X}_{:,m} \tag{5.2.6}$$
$$m \in \{1 ... M\}$$

となる．$\boldsymbol{\theta} \in \mathbb{R}^{K \times 2}$ はパラメータとなり、$\boldsymbol{D_O^{-1}} \boldsymbol{W}$ と $\boldsymbol{D_I^{-1}} \boldsymbol{W}^T$ は出次と入次を考慮した拡散過程の遷移行列を表す．式 (5.2.6) で定義された Convolution を用いて入力の M 次元から出力 Q 次元へと写像する Diffusion Convolutional Layer は出

力 $H \in \mathbb{R}^{n \times Q}$ のとき，

$$\boldsymbol{H}_{:,\boldsymbol{q}} = \boldsymbol{a}\left(\sum_{m=1}^{M} \boldsymbol{X}_{:,\boldsymbol{m}} {}^{\star}\boldsymbol{g}\boldsymbol{f}_{\boldsymbol{\Theta}_{\boldsymbol{q},\boldsymbol{m},:,:}}\right) \quad q \in \{1...\boldsymbol{Q}\} \tag{5.2.7}$$

と定義される．$\boldsymbol{\Theta} \in \mathbb{R}^{Q \times M \times K \times 2}$ はパラメータ，$\boldsymbol{a}$ は ReLU や Sigmoid などの活性化関数を表す．

さらに，時間依存性を考慮するために，Diffusion Convolutional Layer を RNN に取り組む．本モデルは RNN の 1 つとして Gated Recurrent Neural Networks[51] を用いて行列乗算を Diffusion Convolution に置き換える．

$$\begin{aligned}
\boldsymbol{r}^{(t)} &= \sigma\left(\boldsymbol{\Theta}_r {}^{\star}g\left[\boldsymbol{X}^{(t)}, \boldsymbol{H}^{(t-1)}\right] + \boldsymbol{b}_r\right) \\
\boldsymbol{u}^{(t)} &= \sigma\left(\boldsymbol{\Theta}_u {}^{\star}g\left[\boldsymbol{X}^{(t)}, \boldsymbol{H}^{(t-1)}\right] + \boldsymbol{b}_u\right) \\
\boldsymbol{C}^{(t)} &= tanh\left(\boldsymbol{\Theta}_C {}^{\star}g\left[\boldsymbol{X}^{(t)}, \left(\boldsymbol{r}^{(t)} \odot \boldsymbol{H}^{(t-1)}\right)\right] + \boldsymbol{b}_C\right) \\
\boldsymbol{H}^{(t)} &= \boldsymbol{u}^{t} \odot \boldsymbol{H}^{(t-1)} + \left(1 - \boldsymbol{u}^{(t)}\right) \odot \boldsymbol{C}^{(t)}
\end{aligned}$$

このとき，$H^{(t)}$ は時点 t における出力，$\boldsymbol{r}^{(t)}$，$\boldsymbol{u}^{(t)}$ はそれぞれ時点 t における reset gate, update gate を表す．$\boldsymbol{\Theta}_r, \boldsymbol{\Theta}_u, \boldsymbol{\Theta}_C$ はパラメータである．

2 週以上先の予測のために DCRNN モデルは Encoder-Decoder のアーキテクチャを用い，Decoder 部分は 1 つ前の正解値（テスト時にはモデルによる予測値）をもとに予測値を求める．このとき，学習時とテスト時の入力分布間のズレによる性能低下の可能性を軽減するために Scheduled Sampling [52] を用いる．DCRNN モデルの概要を図 5.3 に示す．

5.3 提案モデル

前節で述べたモデルを地域ごとの感染症予測に適応するために，(1) 通勤・通学データの利用，(2) STNN モデルの拡張を行う．

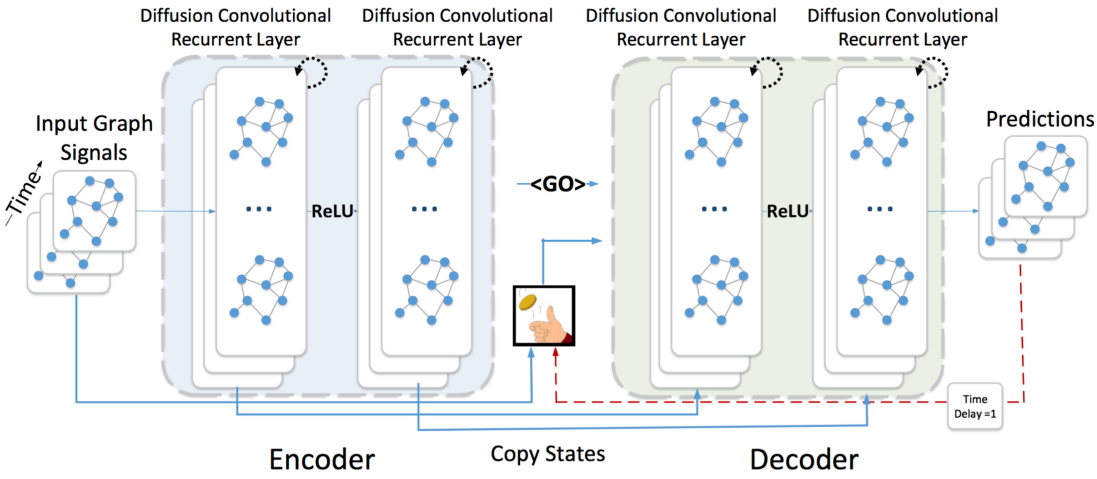


図 5.3　DCRNN モデルの概要 ([50] より引用) 入力系列はエンコーダに入力され，その最終状態がデコーダの初期状態に用いられる．デコーダは 1 つ前の入力値もしくはモデルの出力に基づいて予測を行なう．

5.3.1　通勤・通学データの利用

一般的に交通量予測などに用いられる時空間モデルでは空間を考慮するために観測点間の距離や観測点同士の隣接行列が用いられる．しかし感染症の多くは人から人へと感染していくことから，人の移動経路を考慮することが重要であると考えられる．このことから，本実験では人流を考慮するために，一般的に用いられる地域の隣接情報ではなく，地域間の通勤・通学データを用いる．

5.3.2　STNN モデルの拡張

本章で用いる STNN モデルをより感染症流行の時系列予測問題に適するように，2 つの拡張モデルを提案する．

STNN-RNN (MLP)

インフルエンザなどの感染症の一部は季節性の周期を持つものもあり，感染症患者の時系列には長期の時間依存が見られることもある．一部の感染症はそのような特徴を持つことから，STNN モデルの $g(\boldsymbol{Z_t})$ 関数のように 1 週前の潜在表現から次

の潜在表現を予測することは難しいと考えられる．このことから，1 週前だけでなく，より長期間の潜在表現の時間依存性を捉えられるように RNN もしくは Multi Layer Perceptron 機構の入力とし，先の潜在表現を予測するモデルに置き換えることが感染症の時系列予測において有用であると言える．STNN モデルの Dynamic 部分を時間依存性を考慮できるように以下のように置き換える．

$$g(\boldsymbol{Z_t}) = RNN(\boldsymbol{Z_t}...\boldsymbol{Z_{t-T}}) \qquad \text{もしくは}$$
$$g(\boldsymbol{Z_t}) = MLP(concat(\boldsymbol{Z_t}...\boldsymbol{Z_{t-T}}))$$

このモデルを STNN-RNN (MLP) とする．

STNN-Other

感染症流行予測において，時系列そのものだけでなく，SNS や検索クエリといった外部リソースも有用であり，それらのリソースを組み合わせることで予測精度が向上することは 3 章と 4 章で述べたとおりである．STNN モデルも同様に，他リソースと時系列を組み合わせて将来の予測を行なうモデルへと拡張することを試みる．具体的には，STNN モデルの Dynamic 部分を $g(\boldsymbol{Z_t})$ から $g\left(\boldsymbol{Z_t^+}\right)$ へと変更する．このとき，以下のように定義される．

$$\boldsymbol{W}_t = MLP(\boldsymbol{O}_t) \qquad \boldsymbol{Z}_t^+ = concat([\boldsymbol{Z}_t, \boldsymbol{W}_t])$$

$\boldsymbol{O}_t$ は時点 t における他リソースの値，$\boldsymbol{W}_t$ は $\boldsymbol{O}_t$ の潜在表現である．$\boldsymbol{O}_t$ を MLP を通して潜在表現 $\boldsymbol{W}_t$ に変換し，その潜在表現 $\boldsymbol{W}_t$ と時系列の潜在表現 $\boldsymbol{Z}_t$ を連結した $\boldsymbol{Z}_t^+$ に $g(.)$ 関数を用いることで次の潜在表現 $\boldsymbol{Z}_{t+1}$ を予測するモデルとなっている．STNN-RNN (MLP) と STNN-Other の概要を図 5.4 に示す．

5.4 実験

本節では，日本の都道府県ごとのインフルエンザ罹患者数の予測を行ない，時空間モデルの有効性について検証を行った結果を示す．

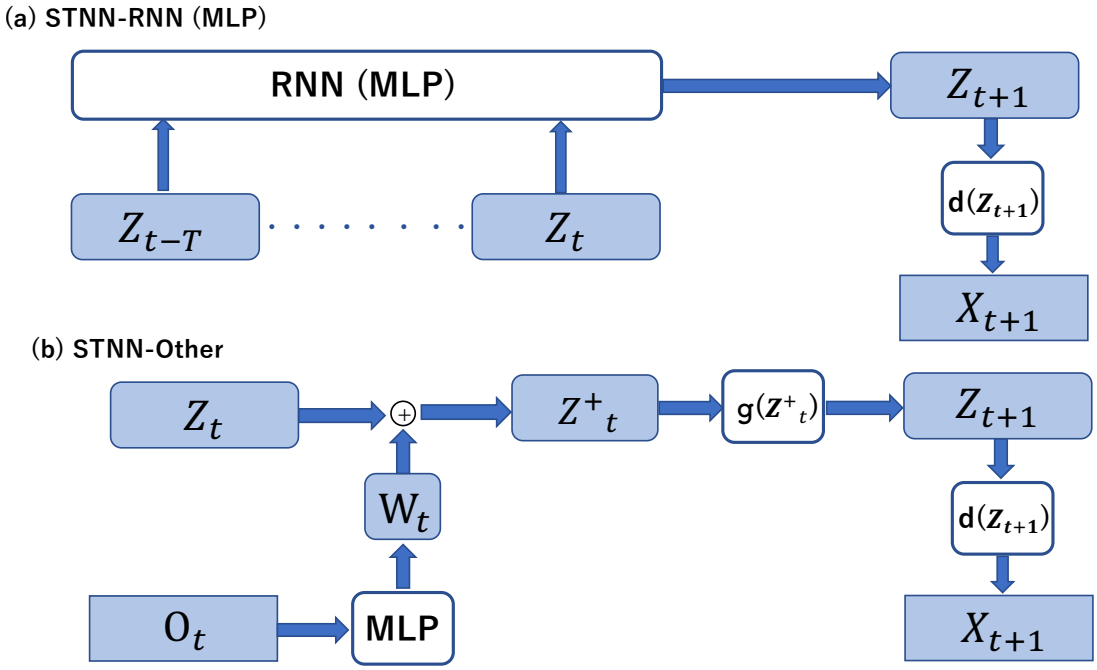


図 5.4　(a)STNN-RNN (MLP)，(b)STNN-Other モデルの概要

5.4.1　データセット

インフルエンザデータ

インフルエンザデータについては，3 章と同様に NIID の提供しているデータを用いる．

通勤・通学データ

日本の都道府県間の人流を考慮するために，都道府県間の通勤・通学の人口データを用いる．都道府県間の通勤・通学データは平成 27 年国勢調査[†]の「従業地・通学地による人口・就業状態等」の項目で集計されており，本実験ではその集計データを利用する．正規化のために，各都道府県の総人口で分割しパーセント単位に置き換えたデータを用いる．

[†]https://www.stat.go.jp/data/kokusei/2015/kekka.html

Google Trend データ

STNN-Other モデルで用いるリソースとして Google Trend[‡]のデータを使用する．本実験では，インフルエンザの流行と関連があると考えられる 5 つのクエリ「インフルエンザ」「インフル」「タミフル」「加湿器」「潜伏期間」を選択する．実験期間中，5 つのクエリの都道府県ごとかつ，1 週間ごとの検索量を取得し，STNN-Other モデルのリソースとして用いる．

5.4.2 設定

時空間モデルを用いて，日本の都道府県ごとのインフルエンザ罹患者数の予測を行なう．本実験では 5 週間先までの予測を行なう．1 週間先の予測は 1.2 節での “Nowcasting”，2 週間先以降の予測は “Forecasting” のことを指す．訓練期間を設けて学習したモデルを用いてテストを行なう静的なモデルとしてベースラインとなる LSTM（3.2.2 節参照），DCRNN モデルの 2 つ，入力データが与えられると都度学習を行なう 4 つの動的なモデルとして STNN，STNN-MLP，STNN-RNN，STNN-Other モデルの計 6 つのモデルで，1 週間先から 5 週間先までの予測を行い時空間モデルの有効性を検討する．STNN-Other モデルは，入力の 1 つとなる他リソースが 2 週間先以降の予測のためのデータを取得できないことから 1 週間先の予測のみ行なう．静的なモデルと動的なモデルでは学習の手法が異なるため単純には精度比較ができないことを注意されたい．

静的なモデルの訓練期間として 2012 年 37 週–2015 年 37 週の 156 週の 3 年間，検証期間として 2015 年 38 週–2016 年 14 週の 30 週，テスト期間として 2016 年 15 週–2018 年 14 週の 104 週の 2 年間を設ける．ベースラインの LSTM は，季節性が見られることから 26 週を入力とし，ハイパーパラメータの隠れ層の次元を 5，20，50，80，150，200 から検証期間を用いて選択する．DCRNN モデルは 26 週を入力，学習率を 0.01 と設定し，ハイパーパラメータの隠れ層の大きさを 32，64，128，256 から検証期間を用いて選択する．動的なモデルも同様に訓練期間を 3 年とし，テスト期間は静的なモデルと同様とする． STNN モデルの潜在表現の次元数 N を 5，λ を 0.1，STNN-RNN(MLP) の T を 26，隠れ層の大きさを 80，STNN-Other

[‡]https://trends.google.co.jp/trends/

の $\boldsymbol{W}_t$ の次元数を 5 と設定する．

5.4.3 評価指標

モデルの予測性能を比較するために 2 つの指標：決定係数 R^2 と平均絶対誤差 MAE を用いる (3.3.2 節参照)．本実験では $MAPE$ を評価指標として用いないが，これは正解値に 0 が含まれることにより正しい評価ができないためである．

5.4.4 結果と考察

実験結果を表 5.1 に示す．DCRNN モデルはベースラインの LSTM モデルと比較して，全体的に精度が向上している．特に，4，5 週先の予測の精度が向上している．このことは，DCRNN モデルに含まれている地域間の人流データと，5 週先予測までを考慮した Encoder-Decoder モデルが良い働きをした結果だと考えられる．また，この結果は GCN が感染症予測において有用であることを示しており，今後も多くの国家間や地域間の感染症予測を行なう際には GCN を用いたモデルを選択肢の 1 つとすることは重要であるといえる．

次に，STNN モデルをベースとした動的なモデルについて，どのモデルも 1 週先，2 週先の予測は他モデルと比較して高い精度を保っている一方で，4 週先以降は DCRNN モデルのほうが高い精度となっている．このことは，STNN モデルが長期間先の予測を前提としていないモデルであり，長期間の予測が不安定になることを示していると考えられる．

本章で提案した STNN-MLP は 1 週先，2 週先において最も高い精度となっており，Dynamic 部分の $g(.)$ が直前の時間依存を考慮するだけでなく長期間の時間依存性を考慮することが重要であることがわかる．しかしながら，同様の時間依存性を考慮した STNN-RNN の精度は通常の STNN モデルと比較しても高くなく，RNN 部分のチューニングなどをより慎重に行なう必要がある．STNN-Other モデルも精度が向上しており，これまでの研究と同様に，地域ごとの予測においても，インフルエンザ罹患者数の時系列だけでなく，検索クエリなどの他リソースを組み合わせることでモデルの予測性能が向上することが確認された．

表 5.1　地域ごとにおけるインフルエンザ罹患者数予測精度

予測対象週	評価指標	Static		Dynamic			
		LSTM	DCRNN	STNN	STNN -MLP	STNN -RNN	STNN -Other
1 週先	MAE	169.68	163.40	154.09	**140.59**	169.92	141.12
	R^2	0.909	0.884	0.923	0.932	0.904	**0.942**
2 週先	MAE	283.72	242.83	264.83	**219.73**	324.12	-
	R^2	0.771	0.738	0.848	**0.849**	324.12	-
3 週先	MAE	364.40	**281.09**	416.13	319.22	494.84	-
	R^2	0.656	0.656	0.848	**0.849**	324.12	-
4 週先	MAE	414.34	**299.72**	575.90	403.55	658.93	-
	R^2	0.588	**0.617**	0.478	0.579	0.280	-
5 週先	MAE	451.08	**325.25**	727.21	794.94	658.93	-
	R^2	0.566	**0.593**	0.297	0.498	0.109	-

5.4.5　通学・通勤データの有効性について

本章では，都道府県の隣接データと比較して通勤・通学データが精度向上に寄与しているかどうかの検証を行なう．これまでの実験では，都道府県の位置関係を考慮するために隣接関係ではなく通勤・通学データを用いた．本実験ではこれまでの STNN モデルと DCRNN モデルを用いて，空間情報データとして通勤・通学データを用いた場合と都道府県の隣接情報を用いた場合の比較を行なう．都道府県の隣接情報を用いたモデルをそれぞれ DCRNN-AD, STNN-AD と呼称する．

結果を表 5.2 に示す．空間情報として都道府県間の隣接データを用いた DCRNN-AD および STNN-AD よりも，人流が考慮されていると考えられる通勤・通学データを用いた DCRNN および STNN モデルの方がよい結果となった．また，1 週先などの近い予測よりも，4，5 週先など遠い予測の方が人流を考慮した効果が高い傾向がみられた．本実験の結果，人から人へと移っていくインフルエンザのような感染症を細かな地域ごとに予測する際，人流を考慮することは有用であるという知見が得られた．

表 5.2　通学・通勤データの有効性の検証

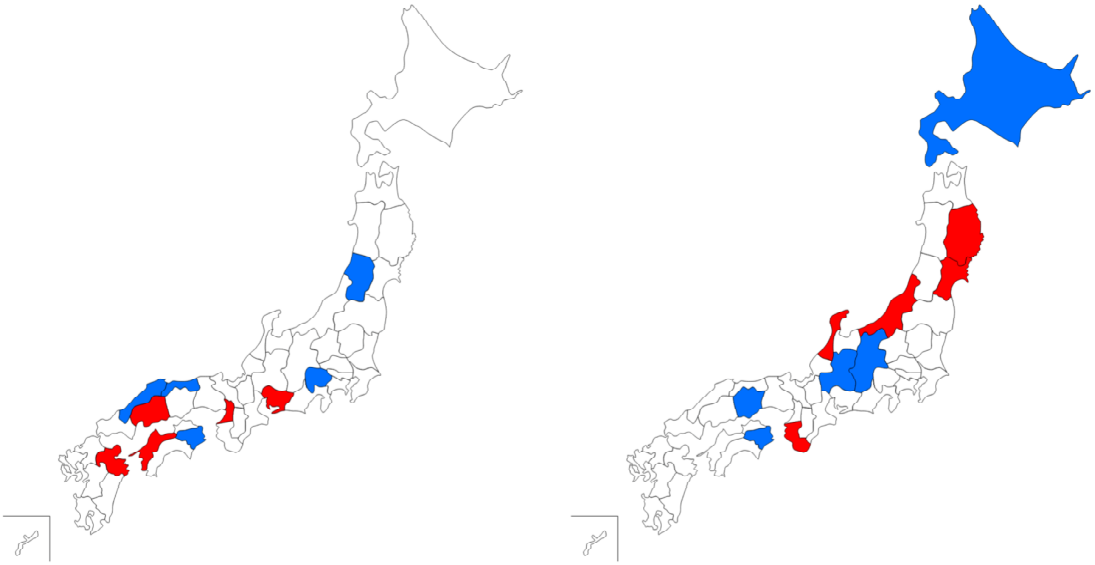
予測対象週	評価指標	DCRNN	DCRNN-AD	STNN	STNN-AD
1 週先	MAE	**163.40**	174.76	**154.09**	155.00
	R^2	**0.884**	0.869	0.923	**0.927**
2 週先	MAE	**242.83**	258.15	**264.43**	275.84
	R^2	**0.738**	0.722	**0.848**	0.827
3 週先	MAE	364.40	281.09	**416.13**	476.99
	R^2	**0.656**	0.601	**0.690**	0.530
4 週先	MAE	**299.72**	342.58	**575.90**	662.06
	R^2	**0.617**	0.504	**0.478**	0.298
5 週先	MAE	**325.25**	389.60	**727.21**	862.38
	R^2	**0.593**	0.418	**0.297**	0.014

5.4.6　DCRNN モデルから検討する時空間モデルの有効性

DCRNN モデルがベースラインの LSTM と比較して精度向上が見られた地域についての検証を行なう．MAE の観点において，LSTM と比較し DCRNN モデルの精度向上が高い 5 都道府県と低い 5 都道府県について表 5.3 に，それらの位置を図 5.5 に示す．これらの結果が示すように，いくつかの地域では MAE において大幅な減少がみられ，インフルエンザ流行予測において高い効果が見られた．このことは，地域間の関係を考慮したことによるものと考えられる．しかしながら，いくつかの都道府県では，DCRNN がベースラインよりも低い精度結果となった．

各都道府県における LSTM と DCRNN モデルの効果の違いはどのような要因で生じているのか．表 5.3 から年度によって，地域の位置関係を考慮することによって精度向上の効果が大きく見られる場所が異なることがわかる．同様に，図 5.5 で都道府県の位置と精度向上の間に関係が無い傾向も見られる．一方で，いくつかの都道府県，例えば，愛知，大分，愛媛，新潟では 2 年とも高い結果，また徳島，岐阜などでは 2 年とも低い結果になっているという傾向も見られる．地域間の関係を考慮することで，精度が向上する地域としない地域が年度によって生じることの要因については今後の課題となるであろう．

最後に，図 5.6 に MAE において 2 年間のテスト期間で最も性能向上が大きい宮城県の時系列の誤差の時系列を示す．この結果が示すように，DCRNN モデルは特



(a) 2016/15 ー 2017/14 週目　　(b) 2017/15 ー 2018/14 週目

図 5.5　MAE の評価指標で LSTM から DCRNN モデルによる予測性能向上の程度が大きい 5 都道府県を赤色，小さい都道府県を青色で示す．

に流行の始まりの時期と終わりの時期に効果が大きい傾向が見られる．

5.5　不確実性の考慮

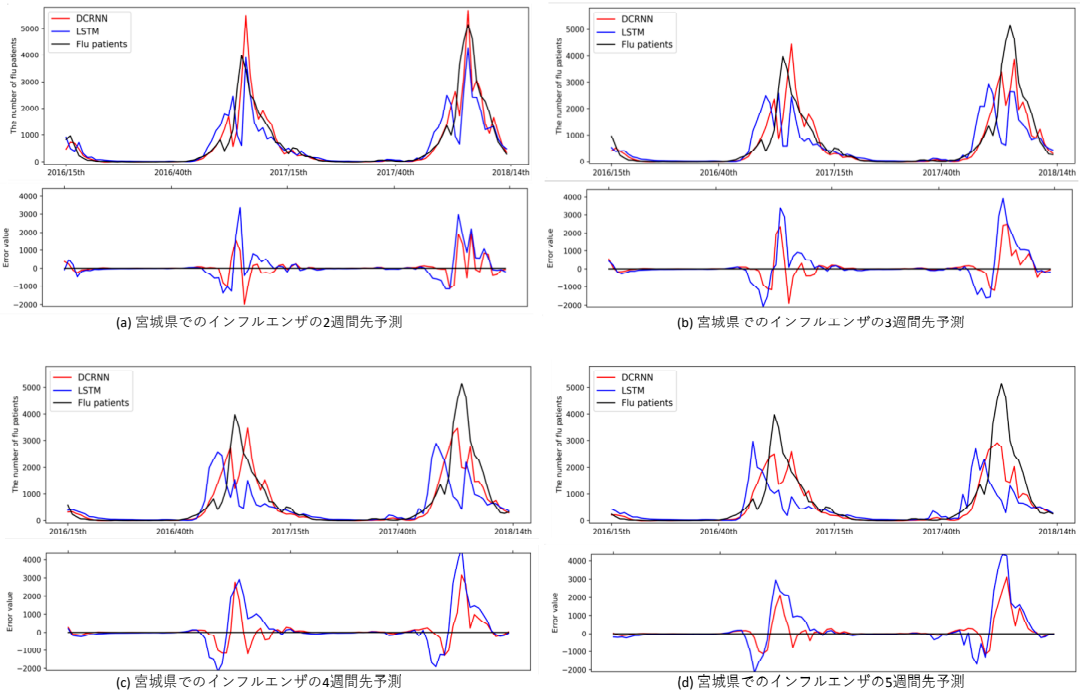
5.5.1　予測区間の必要性

多くの公衆衛生機関にとって感染症の流行は負担の増加やワクチンの需要増加などの問題を引き起こすことが知られている．このような問題を前もって予測できることからも，感染症流行の予測が重要であることはこれまで述べたとおりである．最近は，深層学習を用いた予測モデル [24, 48] も提案されており，本章で用いたモデルもその 1 つである．しかし，深層学習を用いたモデルは点推定であり，予測の上振れや下振れをある程度把握できる予測区間を求めることが困難である．このことは，公衆衛生機関が意思決定を行なう際に，どの程度予測結果を信じればよいのか分からないといった問題に繋がる．

そのような問題に対し，Zhu ら [60] はモデルの不確実性を考慮することで

表 5.3　MAE の評価指標で LSTM から DCRNN モデルによる予測性能向上の程度が大きい 5 都道府県と小さい 5 都道府県．MAE は値が小さいと良い指標であることから，“精度向上の程度” は % のマイナスが大きいほど精度向上していることを示している．2016/15–2017/14 週目における結果時，“もう一方の期間のランク” では 2017/15–2018/14 週目のランクを示す．

2016/15–2017/14 週目			
ランク	都道府県	精度向上の程度 (%)	もう一方の期間のランク
1	愛知	-59.3	12
2	大阪	-57.1	24
3	広島	-57.0	40
4	大分	-56.2	7
5	愛媛	-53.2	13
43	山梨	-1.7	8
44	山形	22.7	9
45	島根	41.0	42
46	徳島	43.8	47
47	鳥取	67.3	39
2017/15–2018/14 週目			
1	秋田	-53.5	33
2	岩手	-49.8	24
3	宮城	-39.6	17
4	新潟	-32.7	10
5	和歌山	-30.4	36
43	北海道	4.8	26
44	長野	7.5	15
45	岡山	17.3	6
46	岐阜	22.9	41
47	徳島	27.7	46



(a) 宮城県でのインフルエンザの2週間先予測

(b) 宮城県でのインフルエンザの3週間先予測

(c) 宮城県でのインフルエンザの4週間先予測

(d) 宮城県でのインフルエンザの5週間先予測

図 5.6　宮城県におけるインフルエンザ流行予測モデルによる DCRNN，LSTM による予測結果，実際の罹患者数の時系列と，それぞれのモデルによるエラーの時系列．それぞれ (a) 2 週先，(b)3 週先，(c)4 週先，(d)5 週先の予測結果を示す．

Encoder-Decoder モデルに対し，予測区間を付与する手法を提案している．この手法をインフルエンザ流行予測の DCRNN モデルに適用した結果を図 5.7(a) に示す．この図から，深層学習のモデルでも予測区間の付与は可能である．しかし，非流行期に罹患者数の変動がほとんど起こらないにも関わらず，予測区間が必要以上に幅が大きい傾向が見られる．実用上において，このような場合は判断に支障をきたす可能性もあり，必要最低限の予測区間を設定することが要求される．そこで，Zhu らの手法に対し，先程の例で見られたような無駄を減らしつつ，予測区間のカバレッジ率を減らさない拡張方法を提案し，その有効性の検証を行なう．

5.5.2 既存手法

予測の不確実性を評価する Zhu らの手法は，予測標準誤差 η の定量化を行い，予測区間を $[y^* - z_{\alpha/2}\eta, y^* + z_{\alpha/2}\eta]$ として求めるものである．

Bayesian neural network では重みパラメータが事前確率として導入され，モデルは最適な事後分布に適用されるように調整される．一般的には W をモデルパラメータとして正規分布 $W \sim \mathrm{N}(0,1)$ が仮定されるとき，

$$y|W \sim \mathrm{N}\left(f^W(x), \sigma^2\right)$$

とデータ生成分布が定義される．このとき，ニューラルネットワークモデルを $f^{\hat{W}}$ とする．N 個の入力セット $X = \{x_1, ..., x_N\}$，$Y = \{y_1, ..., y_N\}$ が与えられたとき，モデルパラメータ W の事後分布 $p(W \mid X, Y)$ の推定を行なう．そして，新しい入力点 x^* に対して事後分布の周辺化を行なうことで，以下の予測分布が得られる．

$$p(y^* \mid x^*) = \int_W p(y^* \mid f^W(x^*)) p(W \mid X, Y)\, dW$$

特に，予測分布の分散は，総分散の定理から

$$\begin{aligned} \mathrm{Var}(y^*|x^*) &= \mathrm{Var}\left[\mathrm{E}(y^*|W, x^*)\right] + \mathrm{E}\left[\mathrm{Var}(y^*|W, x^*)\right] && (5.5.1) \\ &= \mathrm{Var}\left(f^W(x^*)\right) + \sigma^2 && (5.5.2) \end{aligned}$$

となり，モデルの不確実性を表した $\mathrm{Var}\left(f^W(x^*)\right)$ と固有ノイズ部分 σ^2 に分割できる．さらに，5.5.2 節の仮定では，y^* は常に同じ手順で生成されるが，実際は異常パターンの存在などによって大きく異なることもありうる．このことからも，予測の不確実性の推定は，モデルの不確実性，モデルの不完全性，そして固有のノイズの 3 つを考慮し求める必要がある．

モデルの不確実性とモデルの不完全性は，ドロップアウトがベイズと近似可能な性質を利用した MC ドロップアウト [61] を用いることで求めることができる (Algorithm1)．

固有ノイズはホールドアウトされた検証セット $X' = \{x'_1, ..., x'_V\}, Y' = \{y'_1, ..., y'_V\}$ を用いて残差の二乗和から算出を行ない近似する．この固有ノイズは通常よりも過大評価される傾向があり，以下の式で算出される．

$$\hat{\sigma}^2 = \frac{1}{V} \sum_{v=1}^{V} \left(y'_v - f^{\hat{W}}(x'_v)\right)^2 .$$

Algorithm 1 MCdropout

Input: 新しい入力点 x^*, エンコーダ $g(\cdot)$, 予測ネットワーク $h(\cdot)$, ドロップアウト確率 p, イテレーション数 B

Output: 予測結果 $\hat{y}^*_{mc}$, 不確実性 η_1

1: **for** $b = 1$ to B **do**
2: $\quad z^*_{(b)} \leftarrow \mathit{VariationalDropout}(g(x^*), p)$
3: $\quad \hat{y}^*_{(b)} \leftarrow \mathit{Dropout}\ (h(z^*_{(b)}), p)$
4: **end for**
// 予測結果
5: $\hat{y}^*_{mc} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \hat{y}^*_{(b)}$
//モデルの不確実性とモデルの不完全性
6: $\eta_1^2 \leftarrow \frac{1}{B}\sum_{b=1}^{B} (\hat{y}^*_{(b)} - \hat{y}^*)^2$
7: **return** $\hat{y}^*_{mc}$, η_1

モデルの不確実性およびモデルの不完全性と固有ノイズを算出し，不確実性を推定する方法を Algorithm2 に記載する．

5.5.3 提案手法

インフルエンザ流行予測モデルに対し既存手法を用いた際の問題点は，非流行期に罹患者数の変動がほとんど起こらないにも関わらず，予測区間が必要以上に幅が大きい傾向が見られることである．この問題は，既存手法の固有ノイズ推定に起因する．既存手法の固有ノイズは，全期間のどの予測に対しても一定であると仮定して算出されるしかし，インフルエンザの罹患者数のように季節性の周期が存在する時系列の固有ノイズは季節によって大きく異なる．このことから，固有ノイズを算出するために用いる検証セットも同様に季節性を考慮する必要があると考える．そのため，固有ノイズを算出するために，時系列の周期における同じ時期の検証セットのみを用いる．具体的には，時系列における 1 周期（例：1 年間）の検証セットを用意し，新たな入力が 1 月の入力 x^*_m であれば，検証セットの 1 月周辺のウィンドウ幅 W 分のデータ $\{x'_{m-(W-1)/2}, \cdots, x'_{m+(W-1)/2}\}$ を用いて固有ノイ

Algorithm 2 Inference

Input: 新しい入力点 x^*, エンコーダ $g(\cdot)$, 予測ネットワーク $h(\cdot)$, ドロップアウト確率 p, イテレーション数 B

Output: 予測結果 $\hat{y}^*_{mc}$, 不確実性 η

// モデルの不確実性とモデルの不完全性の推定

$\hat{y}^*, \eta_1 \leftarrow MCdropout\ (x^*, g, h, p, B)$

// ノイズの推定

for x'_v **in** validation set $\{x'_1, ..., x'_V\}$ **do**

  $\hat{y'}_v \leftarrow h(g(x'_v))$

end for

$\eta_2^2 \leftarrow \frac{1}{V} \sum_{v=1}^{V} \left(\hat{y'}_v - y'_v\right)^2$ // 不確実性の合算

$\eta \leftarrow \sqrt{\eta_1^2 + \eta_2^2}$

return $\hat{y}^*, \eta$

ズを算出する．1 周期におけるデータ数を M としたとき，1 周期分の検証セット $X' = \{x'_1, ..., x'_M\}, Y' = \{y'_1, ..., y'_M\}$ を用意し，Algorithm2 を Algorithm3 に置き換える．

5.5.4 実験

インフルエンザ罹患者数の予測モデルから予測区間を得る手法として提案手法が有用かどうかを検討する．そのために，4 章で作成した DCRNN によるインフルエンザ罹患者数予測モデルを用いて，2016 年 38 週目–2017 年 37 週目の期間における 95% 予測区間のバンド幅の平均長と正解値のカバレッジ率を比較する．提案手法のために，1 年分の 2015 年 38 週目–2016 年 37 週目までを検証セットとし，ウィンドウ幅 W を 5 と設定する．

Algorithm 3 季節性を考慮したノイズの推定

Input: 入力点 x^*_m，入力点の周期タイミング m，ウィンドウ幅 W

Output: 予測結果 $\hat{y}^*_{mc}$, 不確実性 η

// モデルの不確実性とモデルの不完全性の推定

$\hat{y}^*, \eta_1 \leftarrow MCdropout\ (x^*_m, g, h, p, B)$

// ノイズの推定

for x'_w **in** $\{x'_{m-(W-1)/2}, ..., x'_{m+(W-1)/2}\}$ **do**

  $\hat{y'}_w \leftarrow h(g(x'_w))$

end for

$\eta_2^2 \leftarrow \frac{1}{W}\sum_{w=1}^{W}\left(\hat{y'}_w - y'_w\right)^2$ // 不確実性の合算

$\eta \leftarrow \sqrt{\eta_1^2 + \eta_2^2}$

return $\hat{y}^*, \eta$

5.5.5 予測区間に関する実験結果

結果を表 5.4 に示す．95% 予測区間の平均バンド幅については，既存手法から約 22–32% の減少に成功している．一方，カバレッジ率は約 97% に留まっており，提案手法が効果的であることが示された．図 5.7(a) と同じ期間，地域，モデルに対して，既存手法のを拡張した提案手法を用いて予測区間の設定を行った結果を図 5.7(b) に示す．設定された予測区間が示すように，正解値が予測区間から外れることなく，非流行期の予測区間が狭まっていることが確認できる．本章で提案した予測区間の幅を減らしつつ不確実性を推定する手法は，検証データが 1 周期分必要という欠点も存在する一方で，インフル罹患者数に限らず交通量データや販売量データなどの時間の周期性を持つ時系列に対しては有用であると考えられる．

DCRNN モデルを用いて，提案した予測区間の付与を行った 1–5 週先まで結果を図 5.8–図 5.16 に記載する．

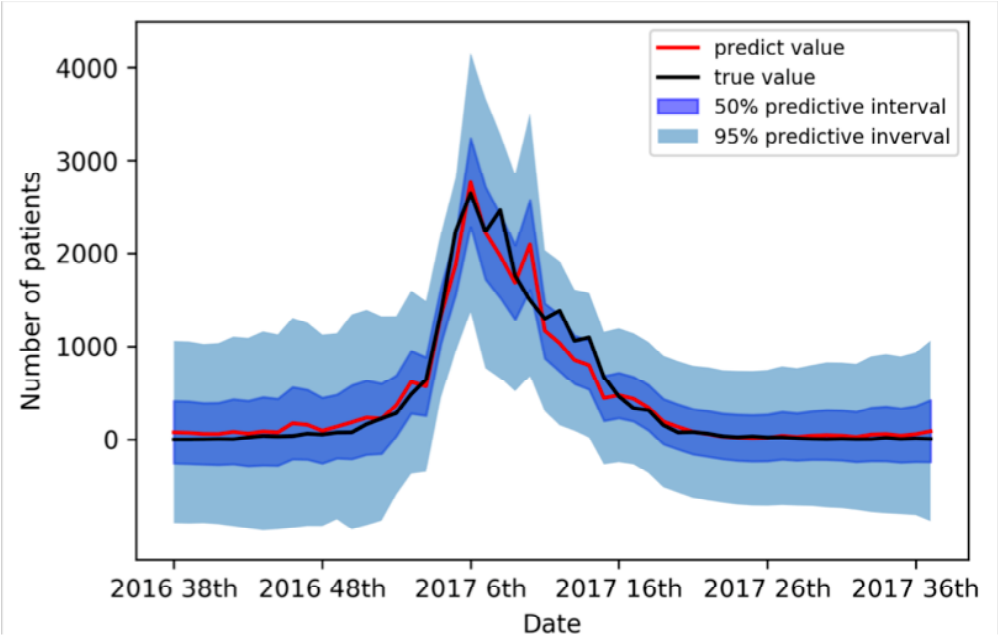
5.6 本章のまとめ

本章では空間情報を考慮するモデルを用いて以下のことを行った．

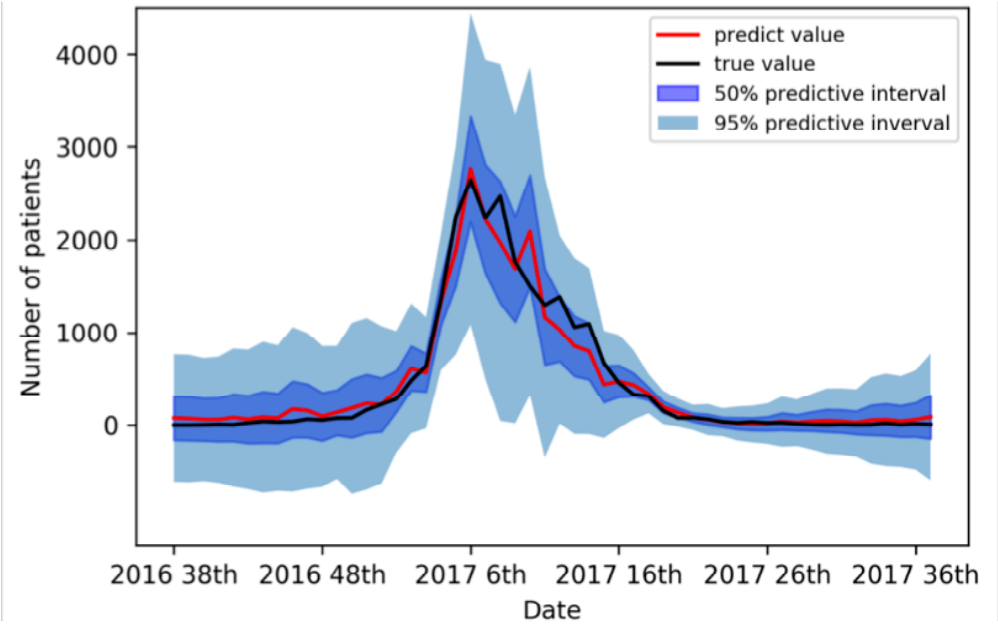
表 5.4　95% 予測区間の平均バンド幅とカバレッジ率

	平均バンド幅		カバレッジ率	
予測対象週	既存手法	提案手法	既存手法	提案手法
1 週先	1687.52	1156.28	99.26%	98.83%
2 週先	2657.21	1978.02	98.64%	97.13%
3 週先	3389.49	2698.25	98.48%	96.52%
4 週先	4042.54	3181.74	98.36%	97.01%
5 週先	4611.41	3299.81	98.36%	97.09%

1. 空間情報として隣接情報ではなく，通勤・通学データを利用することの有効性を検討した
2. Graph Convolutional Network など空間情報を取り組んだ新しいモデルを拡張および利用し，都道府県ごとのインフルエンザ流行予測モデルを作成した
3. 周期性の持つ時系列に対する新たな不確実性推定方法を提案し，深層モデルの予測結果に対して予測区間を付与した



(a) 既存手法



(b) 提案手法

図 5.7　佐賀県での 2016 年 38 週目–2017 年 37 週目における DCRNN モデルを用いた 2 週間先の予測と既存手法を用いた予測区間．(a) は既存手法により推定された予測区間推定で，(b) は提案手法を用いた予測区間推定を表す．薄青の部分は 95% 予測区間，濃青の部分は 50% 予測区間を示す．

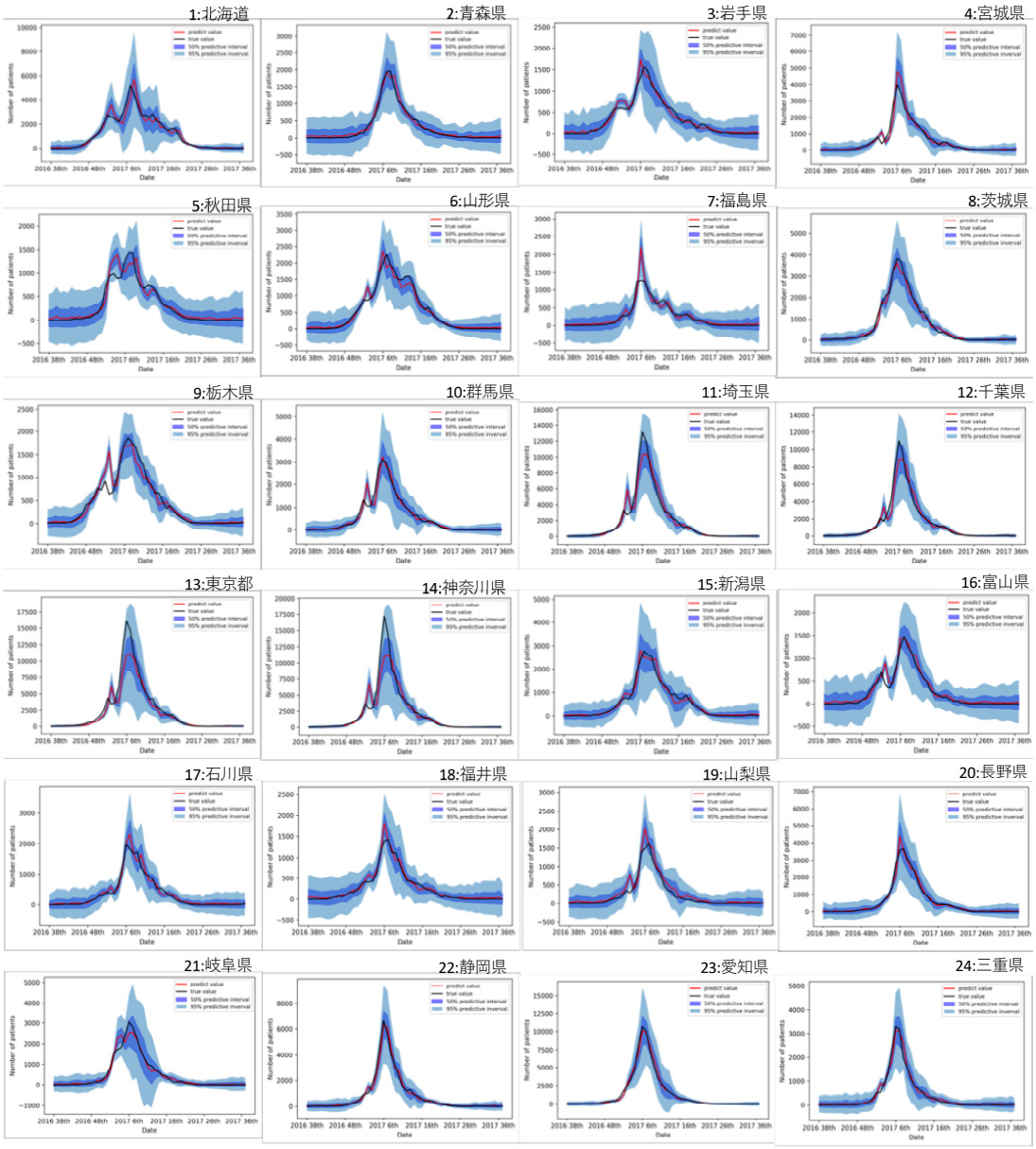


図 5.8　DCRNN による 1 週先予測と提案手法による予測区間の付与：北海道から三重県

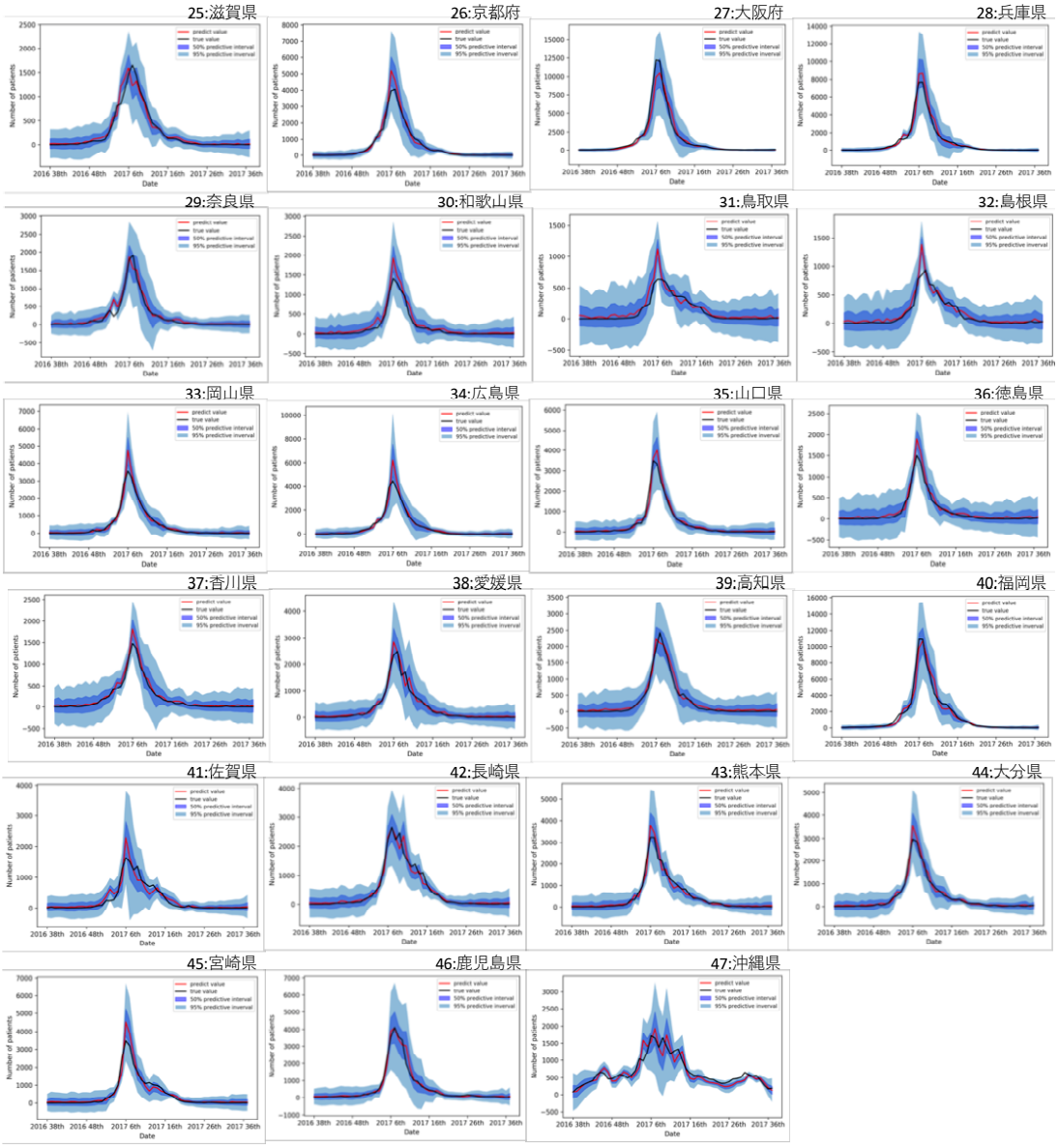


図 5.9　DCRNN による 1 週先予測と提案手法による予測区間の付与：滋賀県から沖縄県

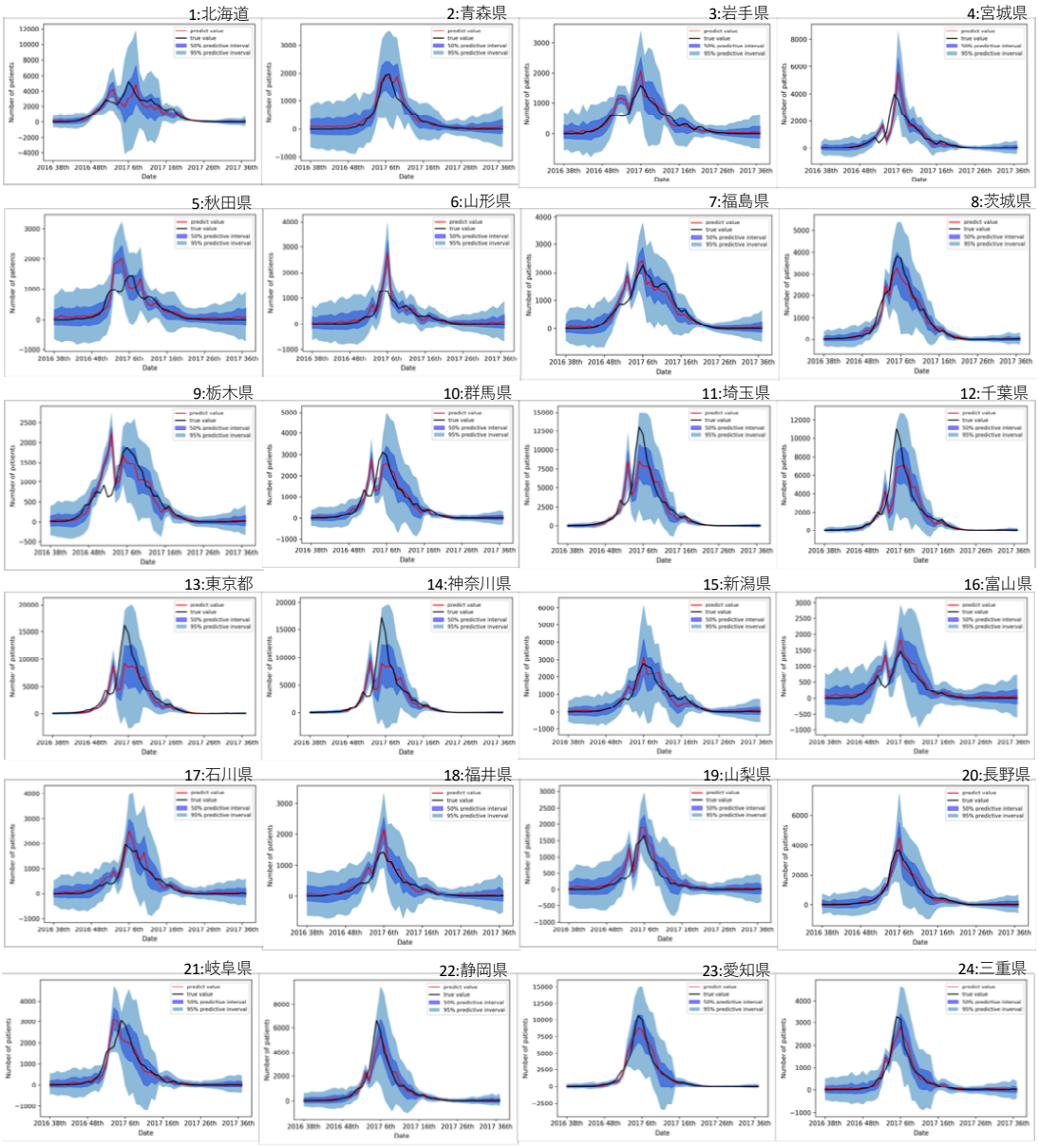


図 5.10　DCRNN による 2 週先予測と提案手法による予測区間の付与：北海道から三重県

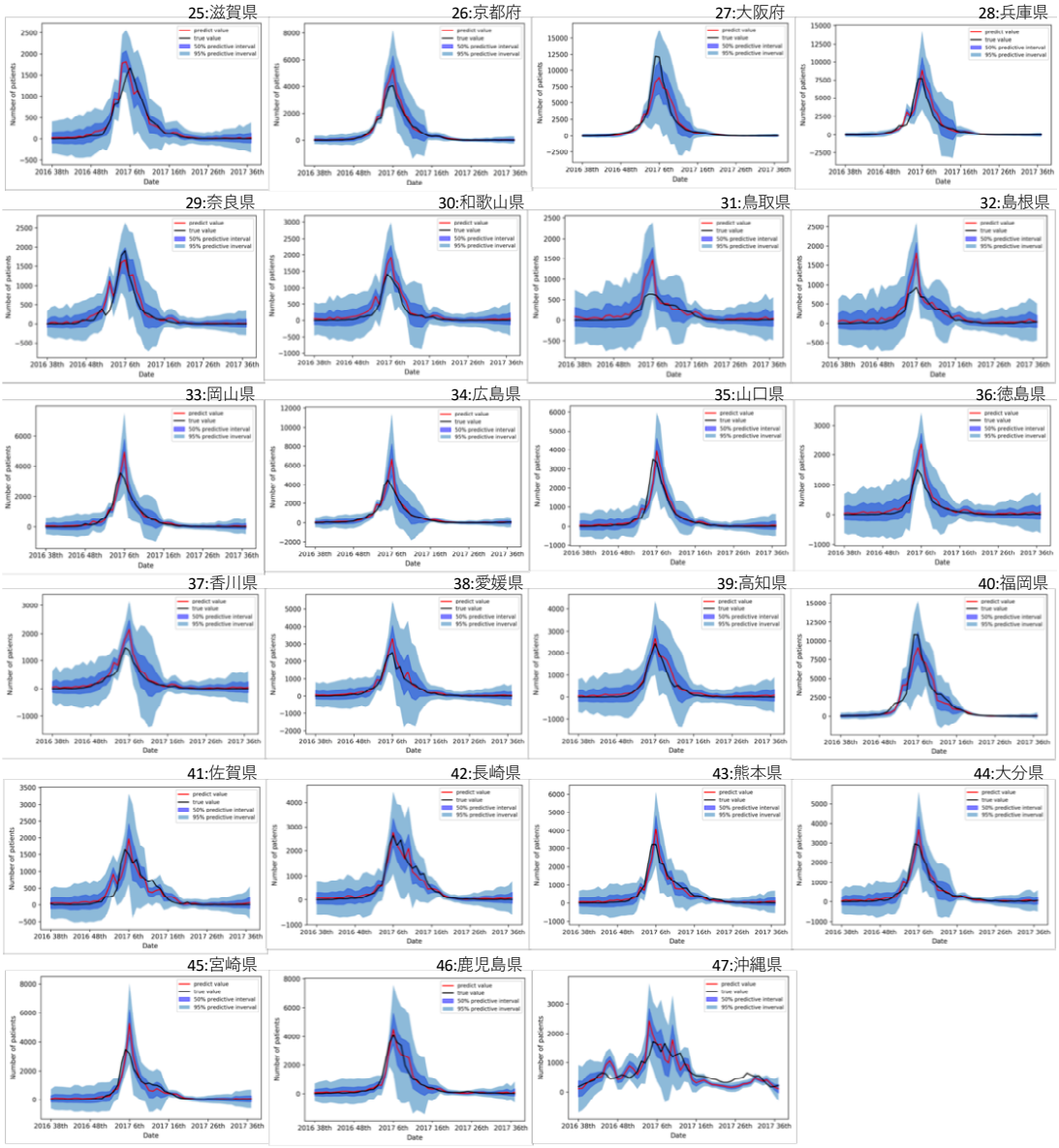


図 5.11　DCRNN による 2 週先予測と提案手法による予測区間の付与：滋賀県から沖縄県

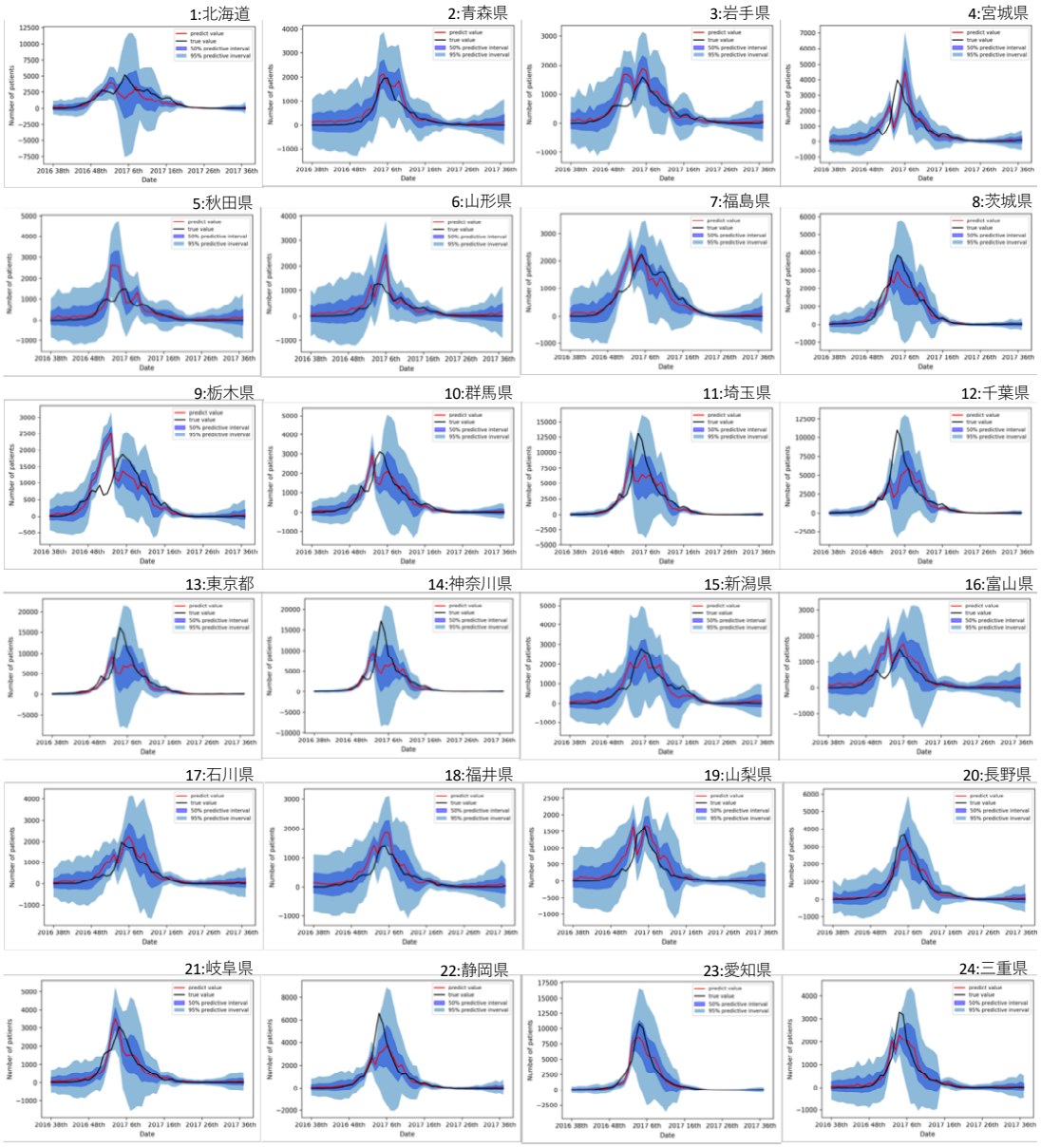


図 5.12　DCRNN による 3 週先予測と提案手法による予測区間の付与：北海道から三重県

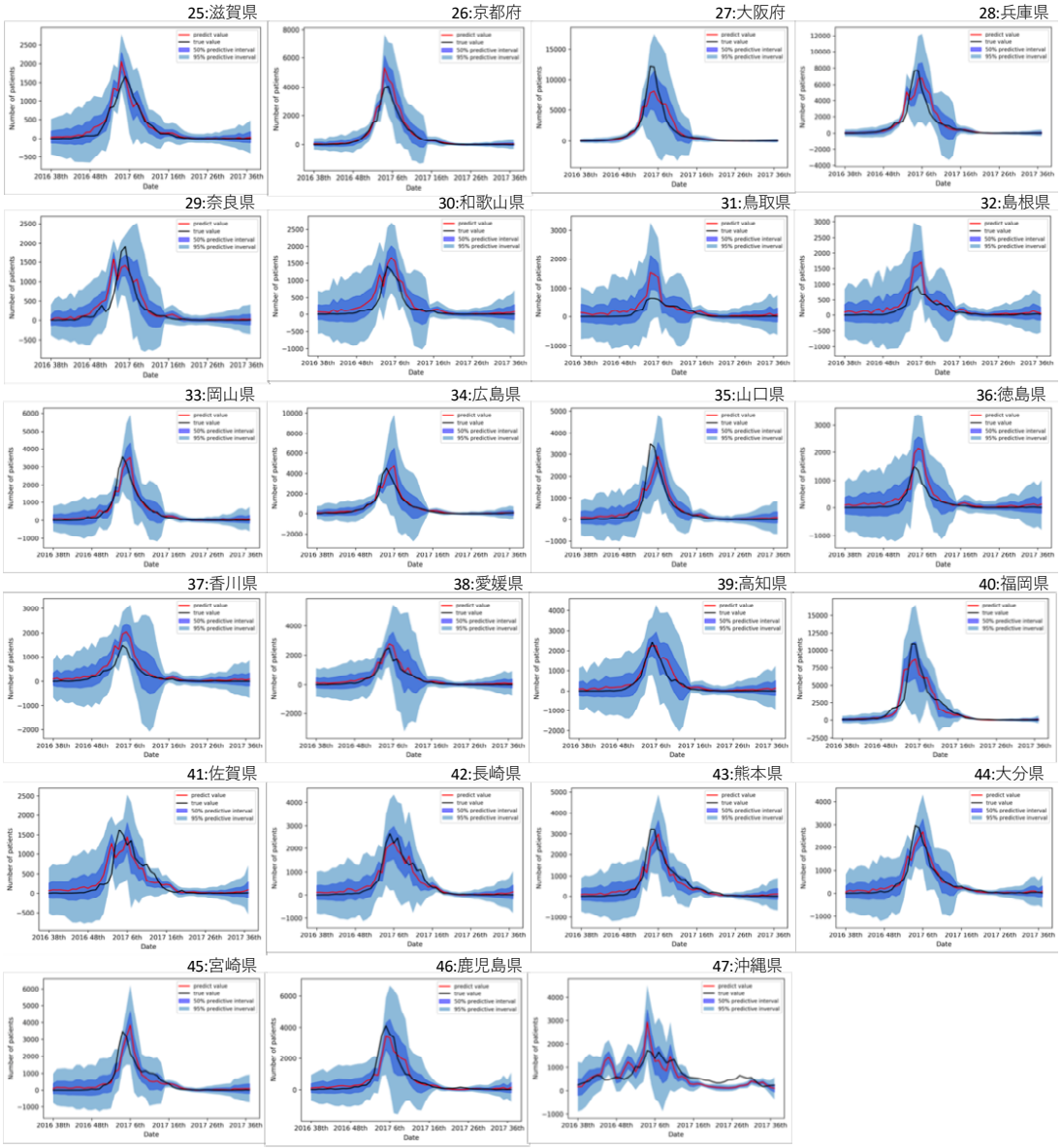


図 5.13　DCRNN による 3 週先予測と提案手法による予測区間の付与：滋賀県から沖縄県

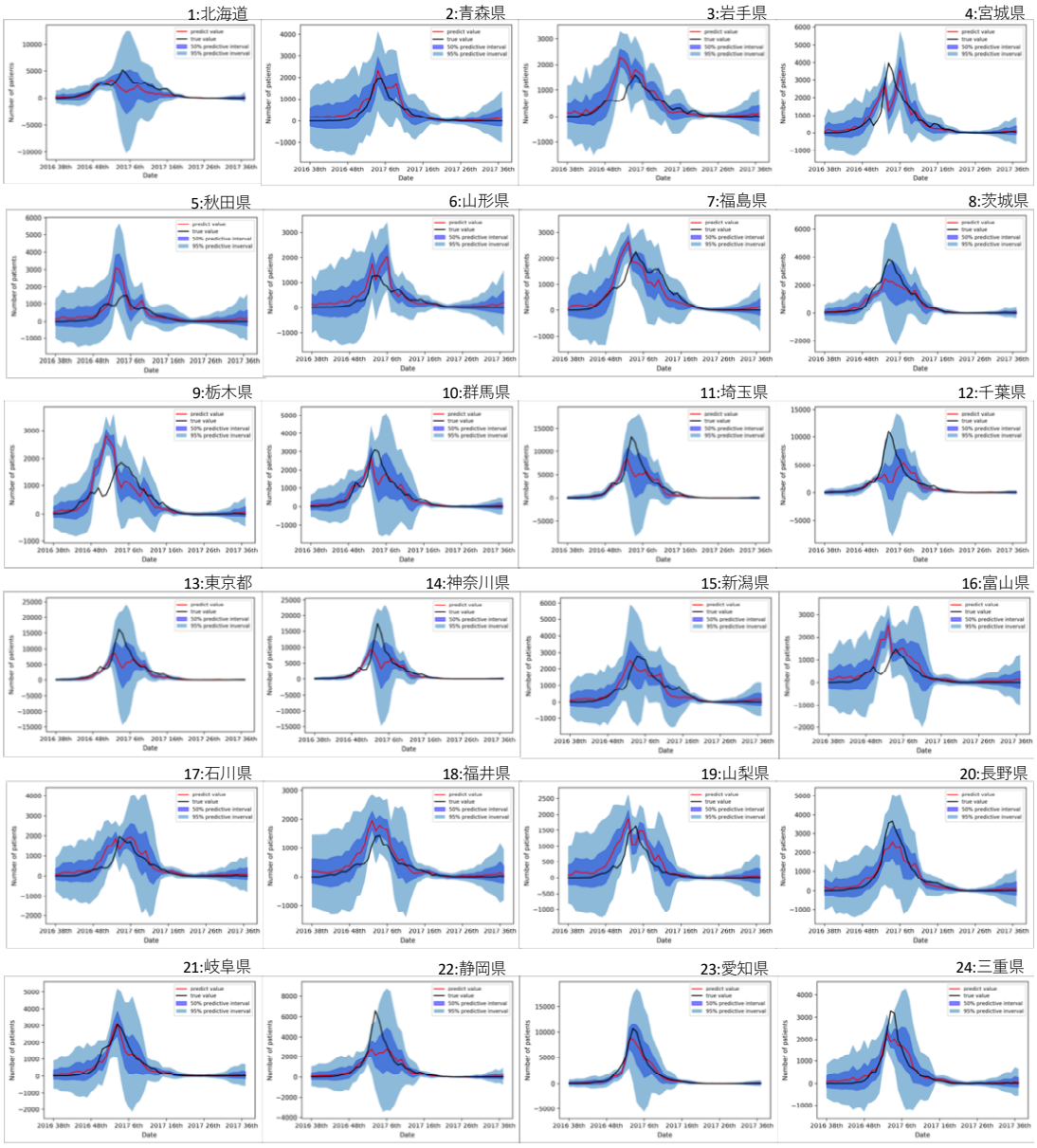


図 5.14　DCRNN による 4 週先予測と提案手法による予測区間の付与：北海道から三重県

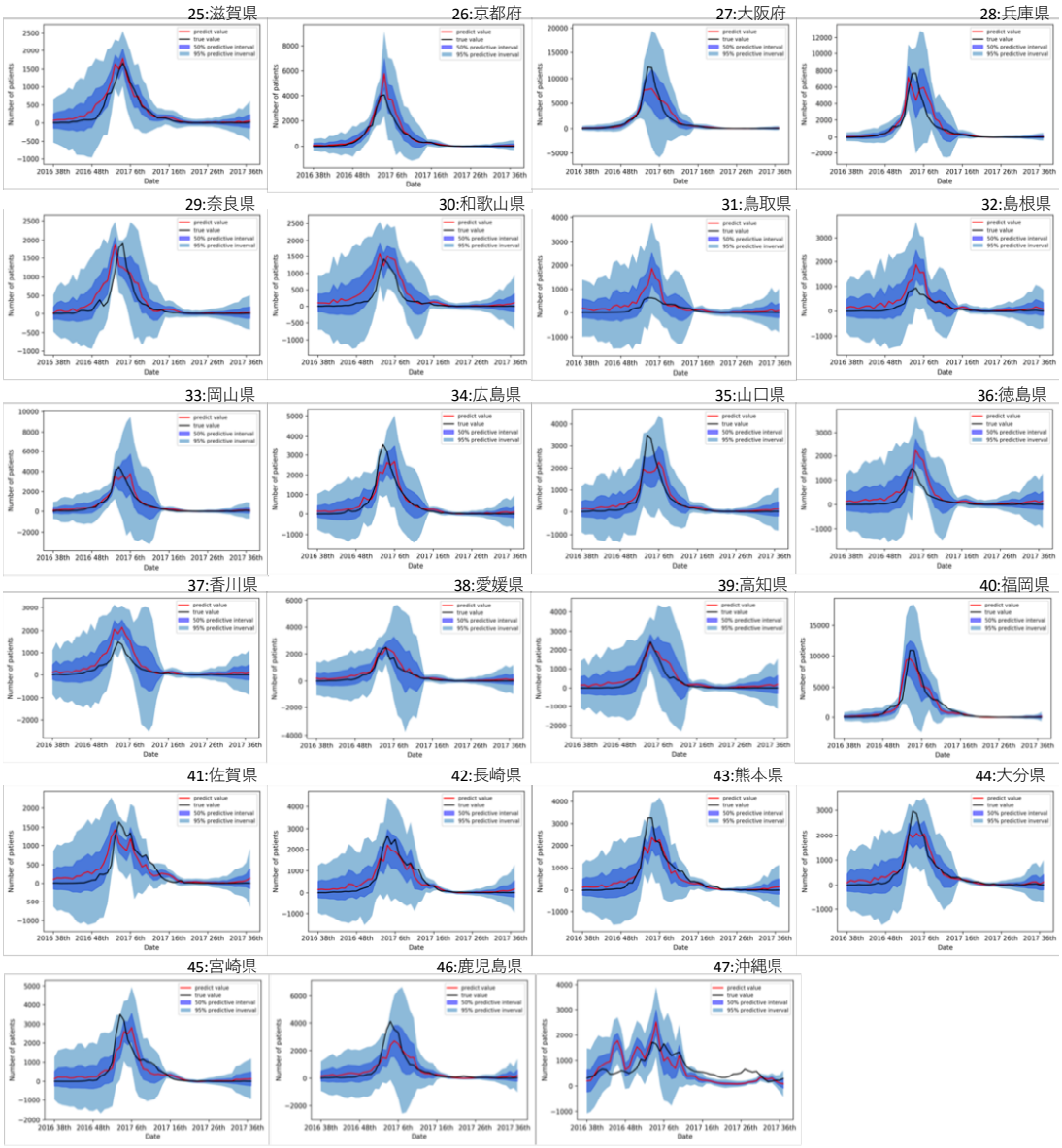


図 5.15　DCRNN による 4 週先予測と提案手法による予測区間の付与：滋賀県から沖縄県

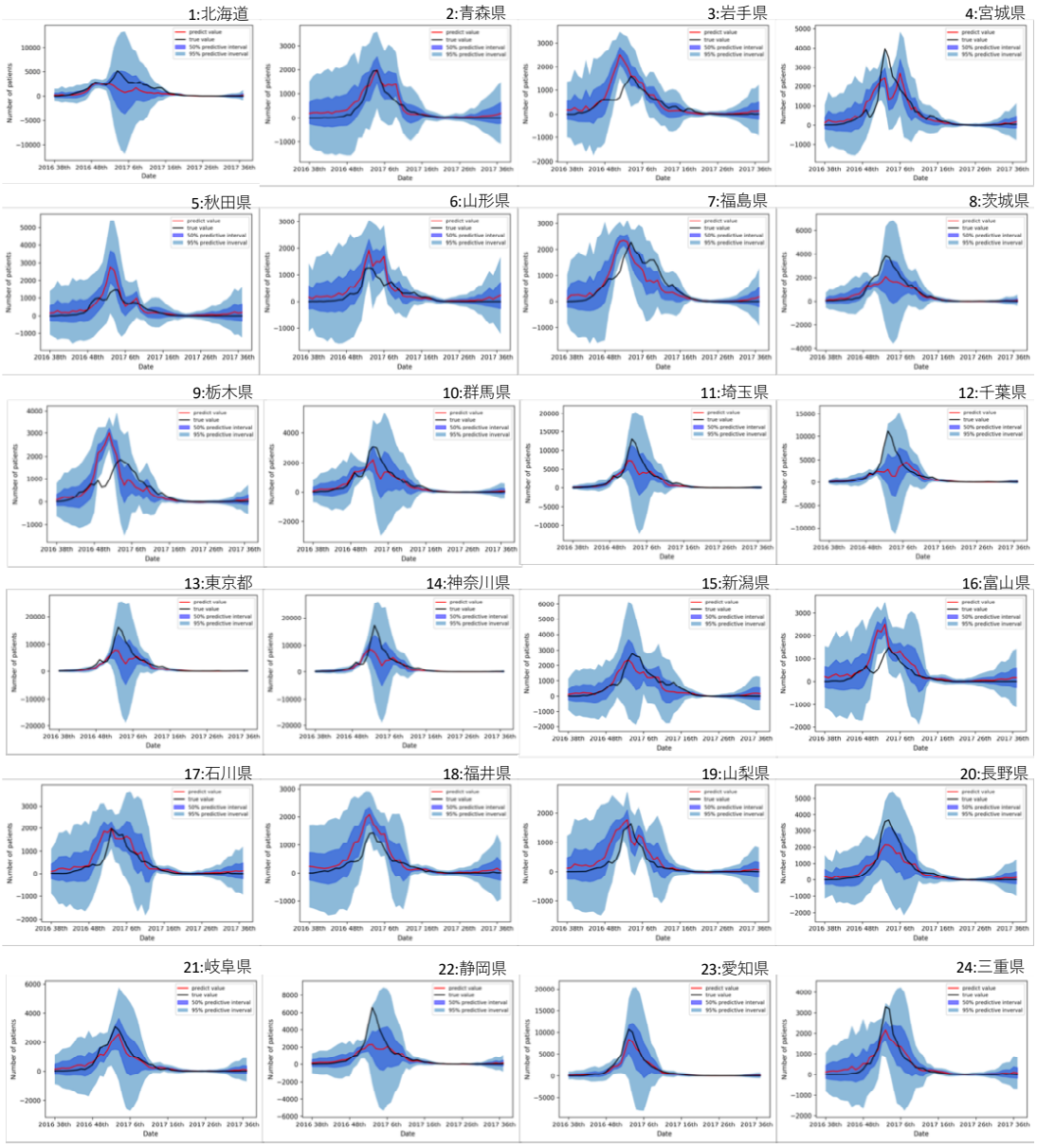


図 5.16　DCRNN による 5 週先予測と提案手法による予測区間の付与：北海道から三重県

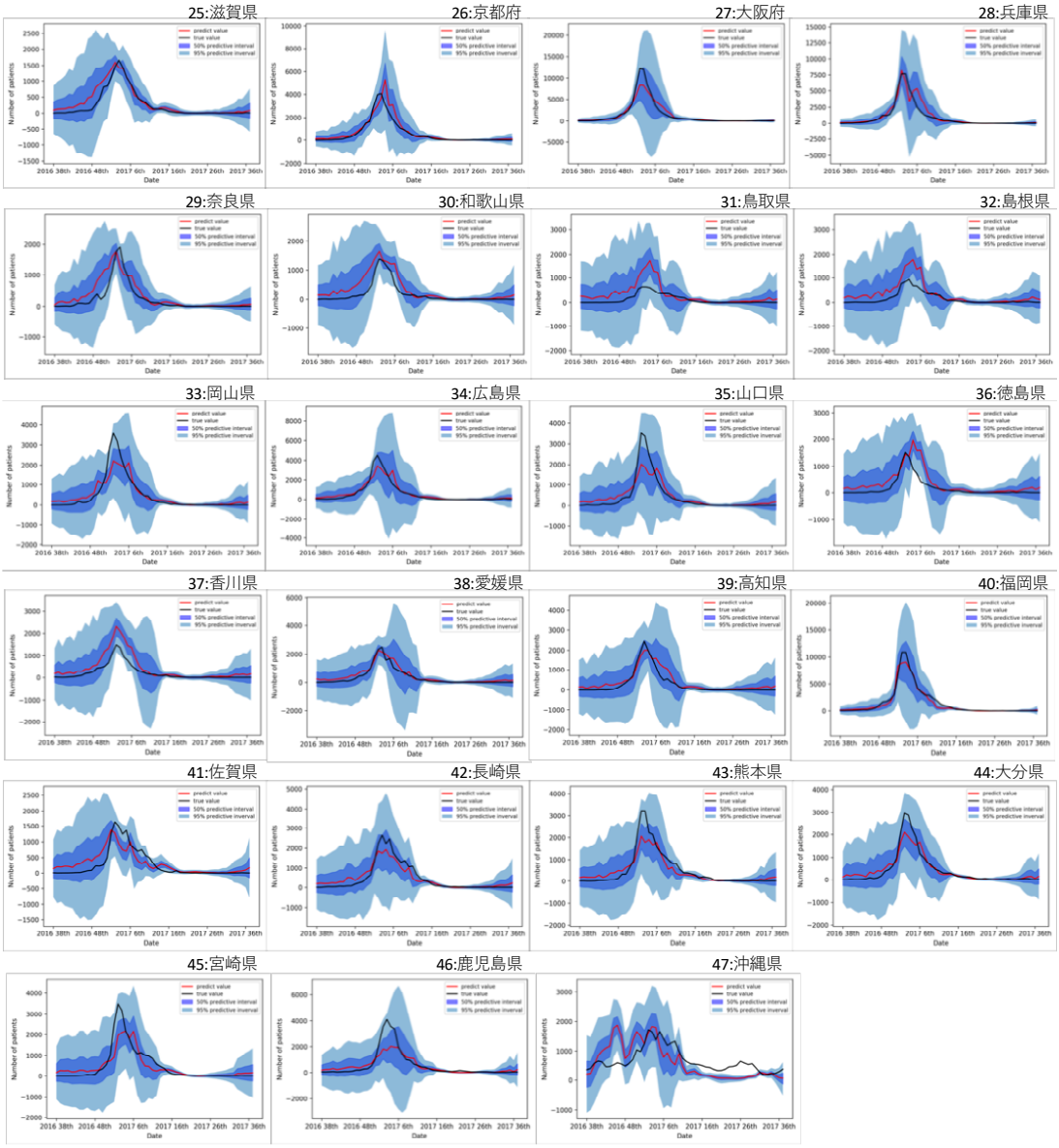


図 5.17　DCRNN による 5 週先予測と提案手法による予測区間の付与：滋賀県から沖縄県

第 6 章　おわりに

本論文では，インフルエンザ流行予測に対して，リソース・モデル・エリアという 3 つの観点から問題を定義し，分析もしくはモデル作成を行い，解決を行った．まず，リソースにおいては，SNS への投稿や検索クエリ，さらにはこれまで用いられることが少なかったオンラインショッピングの検索クエリや Q&A サービスにおける検索クエリの有効性について検証を行った．その結果，特に過去の罹患者数は重要な特徴量であり，SNS への投稿や検索クエリなどの特徴もモデルにおいて有用であった．一方で，Q&A サービスとショッピングの検索クエリの検索量については，変化点検出分析を行なうことで，インフルエンザ流行の時系列の変動を十分に反映しきれていないという結果を得た．

次に，モデルとしては，1 つの UGC と過去の罹患者数を別々のモデルで学習を行なう Two-stage Model の提案を行った．インフルエンザ流行予測のテストにおいて，提案手法がベースラインを上回る結果となりモデルの有効性を示した．また，ロバスト性の検証を行なうことで，Two-stage Model が UGC データがノイズの多い場合でも有効であるという知見を得た．

最後にエリアという観点からも，都道府県単位でのインフルエンザ流行予測を通勤・通学データを用いて地域間を考慮し予測を行った．地域ごとのインフルエンザ流行予測において，空間情報を含める際に地域間の隣接情報よりも通勤・通学データを利用することが有用であるという結果を得た．また，公衆衛生機関にとって決定の助けになる予測区間の推定において，周期性の持つ時系列に適した新たな手法の提案を行い有用性について示した．

本章で行ったインフルエンザ流行予測は，負担軽減や準備という観点において，衛生機関にとって有用であると考えられる．

今後の課題として，現状のインフルエンザ流行予測モデルは早い時期から「いつ流行が生じるのか？」「どのぐらい流行が生じるのか？」といった問題に答えることはできない．過去のパターンを学習する過去の罹患者数を用いたモデルや，直近で流行が起こることを察知する UGC を用いたモデルでは上記の課題を解決することは難しい．このような課題に対して，感染症の流行過程を表した SIR モデル [63] や IDEA モデル [32] などの罹患者個人の行動に焦点を当てた数理モデルや，イン

フルエンザの遺伝子情報から予測を行なうアプローチ [22, 23] を発展させることが有用になってくるであろう.

謝辞

修士論文を提出するにあたり，多くの方々にご協力をいただき，この場を借りてお礼申し上げます．本研究を進めるにあたり，日頃より有益な御助言と丁寧なご指導ご鞭撻を賜りました副指導教員の荒牧英治特任准教授には深く感謝申し上げます．そして，若宮翔子特任助教には，本研究を遂行するあたり，多くの有益な御助言と丁寧なご指導を頂き心より感謝申し上げます．また，ヤフー株式会社の清水伸幸様と藤田澄男様にも，研究テーマの立ち上げ時から数多くの御助言を賜りました．また，主指導教官である松本裕治教授，副指導教官である中村哲教授，鈴木優特任准教授にはご多忙の中にも関わらず，本研究の論文審査委員を引き受けて頂き，心より感謝申し上げます．

大学院生活全般にわたり様々な面で支えていただきました，ソーシャルコンピューティング研究室の皆様に心より感謝を申し上げます．最後に大学院修了まで多大の支援を頂いた家族に深く感謝し，ここに謝辞とさせていただきます．

参考文献

[1] Paul Michael J, et al., “Social monitoring for public health.”, Synthesis Lectures on Information Concepts, Retrieval, and Services, 9.5, 2017.

[2] WHO, Fact sheets Influenza (Seasonal), Available at `http://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)`

[3] Ginsberg et al., Detecting influenza epidemics using search engine query data., Nature, 457(7232):1012-1014, 2009.

[4] Gunther Eysenbach, Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet., J Med Internet Res 11.1, 2009.

[5] Liu Liyuan, et al., “LSTM Recurrent Neural Networks for Influenza Trends Prediction.”, International Symposium on Bioinformatics Research and Applications., Springer, pp.259–264, 2018.

[6] Xu Qinneng et al., Forecasting influenza in Hong Kong with Google search queries and statistical model fusion., PLoS One 12.5, 2017.

[7] Sharpe J et al., Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis., JMIR Public Health and Surveillance 2.2, 2016

[8] Yang Shihao et al., “Accurate estimation of influenza epidemics using Google search data via ARGO.”, In Proc. of the National Academy of Sciences 112.47, 2015.

[9] Lampos Vasileios, et al., “Advances in nowcasting influenza-like illness rates using search query logs.”, Scientific reports, 5, 2015.

[10] Santillana Mauricio et al., Combining search, social media, and traditional data sources to improve influenza surveillance., PLoS Computational Biology 11.10, 2015.

[11] Aramaki Eiji et al., Twitter catches the flu: detecting influenza epidemics using Twitter., In Proc. of EMNLP,pp.1568–1576, 2011.

[12] Paul Michael J, Twitter improves influenza forecasting., PLoS Currents 6, 2014.

[13] Paul Michael J et al., Worldwide Influenza Surveillance through Twitter., AAAI Workshop: WWW and Public Health Intelligence, 2015.

[14] Nagar Ruchit et al., A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives, Journal of medical Internet research., 2014.

[15] Signorini Alessio et al., The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic., PLoS One 6.5, 2011.

[16] Lowen Anice C et al., Roles of humidity and temperature in shaping influenza seasonality., Journal of Virology, 2014.

[17] Wu Hongyan et al., Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression., Bioscience Trends 11.3, 2017.

[18] Paolotti D et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection 20.1, 2014.

[19] Jeremy U Espino et al. Telephone triage: a timely data source for surveillance of influenza-like diseases. In AMIA annual Symposium Proceedings, volume 2003, page 215-9, 2003.

[20] S Magruder Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins University APL Technical Digest, 24(4):349–353, 2003.

[21] 梅本 和俊 他. 大規模レセプトデータからの投薬トレンドの変化検知. 第 10 回データ工学と情報マネジメントに関するフォーラム, 2018

[22] Nyirenda, Mayumbo, et al. Estimating the lineage dynamics of human influenza B viruses. PloS one 11.11, 2016.

[23] Kim Kiyeon et al. Host-specific and segment-specific evolutionary dynamics of avian and human influenza A viruses: a systematic review. PloS one 11.1,

2016.

[24] Zhang Jie et al. A comparative study on predicting influenza outbreaks Bioscience Trends 11.5, 533-541, 2017

[25] Dalum Hansenm et al. Seasonal Web Search Query Selection for Influenza-Like Illness (ILI) Estimation. In Proc. of SIGIR Conference on Research and Development in Information Retrieval, 2017.

[26] Lampos Vasileios et al. Enhancing feature selection using word embeddings: The case of flu surveillance. In Proc. of the Web Conference, 2017.

[27] Shin Kanouchi et al. Who caught a cold ? - identifying the subject of a symptom. In Proc. of the Association for Computer Linguistics (ACL), 2015.

[28] Xiao SUN et al. Real time early-stage influenza detection with emotion factors from sina microblog. In Proc. of the Workshop on South and Southeast Asian NLP in the International Conference on Computational Linguistics (COLING), 2014

[29] Zou Bin et al. Multi-Task Learning Improves Disease Models from Web Search. In Proc. of the Web Conference, 2018.

[30] Kane Michael J et al. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 15.1, 2014.

[31] Biggerstaff Matthew et al. Results from the Centers for Disease Control and Prevention predict the 2013–2014 Influenza Season Challenge. BMC Infectious Diseases 16.1, 2016

[32] Nasserie Tahmina et al. Seasonal Influenza Forecasting in Real Time Using the Incidence Decay With Exponential Adjustment Model. Open Forum Infectious Diseases 4.3, 2017

[33] Evan L Ray et al. Prediction of infectious disease epidemics via weighted density ensembles. PLOS Computational Biology, 14.2, 2018.

[34] Dugas, Andrea et al. “Influenza forecasting with Google flu trends”, PLOS one, 8.2, 2013.

[35] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58:267–288, 1994.

[36] Huber Peter J. Robust estimation of a location parameter. The annals of mathematical statistics 35.1:73-10, 1964.

[37] Smola Alex J. "A tutorial on support vector regression." Statistics and computing 14.3:199-222, 2004.

[38] L. Breiman. "Random forests." Machine Learning, vol. 45, no. 1, pp.5—32, 2001.

[39] Barry, Daniel and John A. Hartigan. "A Bayesian analysis for change point problems." Journal of the American Statistical Association 88.421, pp.309–319, 1993.

[40] Erdman Chandra, and John W. Emerson. "A fast Bayesian change point analysis for the segmentation of microarray data." Bioinformatics, 24.19, pp.2143–2148, 2008.

[41] Erdman, Chandra, and John W. Emerson. "bcp: an R package for performing a Bayesian analysis of change point problems." Journal of Statistical Software, 23.3, pp.1–13, 2007.

[42] Sutskever Ilya et al., "Sequence to sequence learning with neural networks" In Proc. of NIPS, pp.3104–3112, 2014.

[43] Hochreiter Sepp et al., "Long short-term memory", Neural Computation 9.9, pp.1735–1780, 1997.

[44] Box George EP, et al., "Time series analysis: Forecasting and control", Holden–Day series in time series analysis, 1976.

[45] Rothman Kenneth J., et al, "Modern epidemiology" ,2008.

[46] Reich Nicholas G., et al., "Forecasting seasonal influenza in the US: A collaborative multi-year, multi-model assessment of forecast performance", bioRxiv, 397190, 2018.

[47] Senanayake R. et al., "Predicting Spatio-Temporal Propagation of Seasonal Influenza Using Variational Gaussian Process Regression", In Proc. of AAAI, pp.3901–3907, 2016.

[48] Wu Yuexin, et al., “Deep Learning for Epidemiological Predictions”, In Proc. of SIGIR, pp.1085–1088, 2018.

[49] Ziat Ali, et al., “Spatio-Temporal Neural Networks for Space-Time Series Forecasting and Relations Discovery”, In Proc. of ICDM, pp.705–714, 2017.

[50] Yaguang Li, et al., “Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting”, In Proc. of ICLR, 2018.

[51] Junyoung Chung, et al., “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv preprint arXiv:1412.3555, 2014.

[52] Samy Bengio, et al., “Scheduled sampling for sequence prediction with recurrent neural networks”, In Proc. of NIPS, pp.1171–1179, 2015.

[53] Bruna, Joan, et al., “Spectral networks and locally connected networks on graphs”, arXiv preprint arXiv:1312.6203, 2013.

[54] Peng Hao, et al., “Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN”, In Proc. of Web Conf., pp.1063–1072, 2018.

[55] Wang Xiaolong, et al., “Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs”, In Proc. of CVPR, 2018.

[56] Duvenaud David K, et al., “Convolutional networks on graphs for learning molecular fingerprints” , In Proc. of NIPS, 2015.

[57] Chai Di, et al., “Bike flow prediction with multi-graph convolutional networks”, In Proc. of SIGSPATIAL, pp.397–400, 2018.

[58] X Geng, et al., “Spatiotemporal Multi-Graph Convolution Network for Ride-hailing Demand Forecasting”, In Proc. of AAAI., 2018.

[59] Yu Bing, et al., “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting”, In Proc. of IJCAI, 2018.

[60] Zhu Lingxue, et al., “Deep and confident prediction for time series at uber”, In Proc. of ICDMW, 2017.

[61] Y. Gal, “Uncertainty in deep learning,” , Ph.D. dissertation, PhD thesis, University of Cambridge, 2016.

[62] Box, George EP, et al., “Time series analysis: Forecasting and control”,

1976.

[63] M. J. Keeling, et al., "Modeling infectious diseases in humans and animals.", Princeton University Press, 2008.

[64] CDC, "FluSight: Flu Forecasting",Available at `https://www.cdc.gov/flu/weekly/flusight/index.html`, 2019.