# インフルエンザ感染者数の傾向分析と予測

9 班　雨谷健司　川﨑航太　早瀬悠希　楊明達

（指導教員 イリチュ美佳）

## 1 研究背景

### 1.1 インフルエンザとは

インフルエンザとは，インフルエンザウイルスによって引き起こされる呼吸器感染症のことである［1］．インフルエンザウイルスにはA，B，C型の3型があり，流行的な広がりを見せるのはA型とB型である．インフルエンザの感染経路は，飛沫感染と接触感染がある．飛沫感染とは，感染した人が咳をすることで飛んだ，飛沫に含まれるウイルスを，別の人が口や鼻から吸い込んでしまい，ウイルスが体内に入り込み感染することである．接触感染とは，感染した人が振り撒いた飛沫を手で触れ，その手で鼻や口に再び触れることにより，粘膜などを通じてウイルスが体内に入り間接的に感染することである．潜伏期間は一般的に1日から3日程度であり，その後高熱が出るのが特徴である．　その体温は 38℃から40℃で，関節痛，背中の痛み，全身の倦怠感，悪寒，頭痛などの症状が全身に出て，鼻水や喉の痛みなどが続く．　完治は発症後5日を経過し，かつ，解熱後2日を経過するまでと言われている．　高齢者，基礎疾患を持つ者，妊婦，乳幼児がインフルエンザにかかると，気管支炎，肺炎などを併発し重症化し，最悪の場合は死に至ることもある．インフルエンザは予防接種により重症化を防ぐことができる．インフルエンザウイルスは常に変異と増殖を繰り返して，徐々にマイナーチェンジしながら生き延びている．そのため，一度感染して免疫を獲得しても別の変異したウイルスに感染してしまう．　さらに，数年から数十年単位でフルモデルチェンジが起こり，世界的大流行を引き起こす．　それゆえに，インフルエンザは，未だ人類に残されている最大級の疫病と言われている．

### 1.2 日本におけるインフルエンザ

インフルエンザウイルスは低温･乾燥を好むため，日本では冬(12月から3月)に流行する．　これは乾燥した冷たい空気より，喉や鼻の粘膜が弱まっていること，年末年始の人の移動でウイルスが全国的に広がることなども原因と言われている．　2009 年には新型インフルエンザが世界的に猛威をふるい，日本にも甚大な被害をもたらした．国立感染症研究所[1]の「第 10-1 表．報告数，週・都道府県・週報定点把握対象疾患・性別」のデータより，2001 年 1 月から 2015 年 12 月までの期間における週ごとの全国インフルエンザ感染者報告数を図 1 に示す．　ここでは，2001 年 1 月第 1 週を 1 としている．

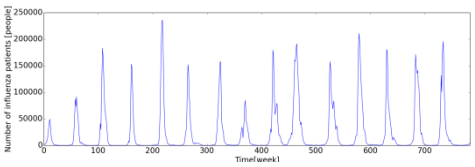


図 1 全国のインフルエンザ患者数の推移

### 1.3 2017 年におけるインフルエンザの流行

インフルエンザワクチンは，他の地域での流行状況やここ数年での傾向を踏まえて，今年はどのような型のウイルスが流行するかを厚生労働省［2］や国立感染症研究所［1］が予測し，毎年異なるものを製造している．　2017/2018 冬シーズンのインフルエンザワクチン株は表 1 のような予想がされている．

表 1　2017/2018 のインフルエンザワクチン株

| ウイルスの型 | 分離された地域 | ウイルス株の番号 | 分離された年 |
|---|---|---|---|
| A | シンガポール | GP1908 | 2015 |
| A | 香港 | 4801 | 2014 |
| B | プーケット | 3073 | 2013 |
| B | テキサス | 2 | 2013 |

## 2 分析・予測手法

本研究では，分析，予測に当たって国立感染症研

究所［1］にて公開されている「第 10-1 表．報告数，週・都道府県・週報定点把握対象疾患・性別」のデータを使用する(図 2)．また，関連する気象データについては，気象庁ウェブサイトよりダウンロードし，使用する［3］．

| | 総　数(total No.) | 1週(week 1) | 2週(week 2) |
|---|---|---|---|
| | 報告数(No. of cases) | 報告数(No. of cases) | 報告数(No. of cases) |
| 総　数(total No.) | 1,169,041 | 100,777 | 164,796 |
| 北海道(Hokkaido) | 43,572 | 5,840 | 4,605 |
| 青森県(Aomori) | 14,893 | 1,467 | 1,786 |
| 岩手県(Iwate) | 15,556 | 2,228 | 1,998 |
| 宮城県(Miyagi) | 19,840 | 1,140 | 2,760 |
| 秋田県(Akita) | 14,387 | 2,261 | 1,371 |
| 山形県(Yamagata) | 10,141 | 471 | 1,008 |

図 2　第 10-1 表(一部抜粋)

### 2.1 季節性の確認

#### 2.1.1 スペクトル解析

時系列データの解析手法の一つにスペクトル解析［4］がある．スペクトル解析とは，不規則なデータを構成周波数成分に分解し，各周波数とエネルギー(振幅)との関係(スペクトル)を取り出すための手法である．そのため，繰り返している現象に有用である．

本研究では，全国と各県インンフルエンザ感染者報告数において 2001 年 1 月から 2015 年 12 月までの週ごとのデータにスペクトル解析を行った(図 3)．

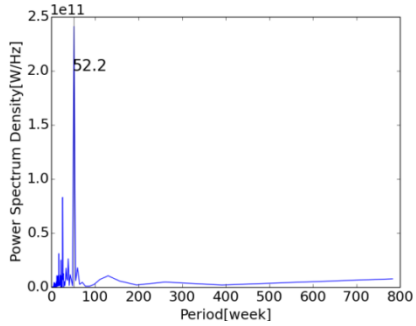


図 3 全国のスペクトル解析結果

図 3 より，全国のインンフルエンザ感染者報告数においての支配的な周期は 52.2 週(約 1 年)であることが読み取れる．また，各県でもスペクトル解析を行ったところ，同様の結果が得られた．このことから，インフルエンザ感染者報告数には季節性があり，気象(温度，湿度，降水量など)との関係性があると考えられる．

#### 2.1.2 相関分析

相関分析［5］とは，2 つ以上の変量の間で，一方の変量が変化すると，他方もそれに応じて変化する関係(相関関係)を統計的に分析することである．その際に相関係数を用いる．

本研究では，気象データと各県のインフルエンザ感染者報告数についての相関分析を行った．さらに，インフルエンザ感染者報告数の前週以前のデータとの相関分析も行った．以上の結果を表 2　茨城県の流行期(2015)と時系列の相関係数にまとめて示す．ここでの流行期には，2001 年 1 月第 1 週を 1 とした時，725 から 750 週目の 2015 年にあたるデータを使用した．

表 2　茨城県の流行期(2015)と時系列の相関係数

| | 時系列 | 流行期(2015) |
|---|---|---|
| 相対湿度 | -0.422 | -0.396 |
| 絶対湿度 | -0.459 | -0.638 |
| 温度 | -0.507 | -0.656 |
| 前週 | 0.941 | 0.935 |
| 前々週 | 0.805 | 0.769 |
| 3 週前 | 0.639 | 0.542 |

使用した気象データは常に変動し続けているが，インフルエンザ感染者報告数は下限(0 人)がある．そのため，時系列データの方は，感染者報告数が下限(0 人)の時に，相関係数に気象データの変動をうまく考慮できていない可能性がある．ゆえに，気象データと感染者報告数との相関は流行期のみの方が相関の高い傾向にあると考えられる．また，一般には，インフルエンザ感染者報告数に最も影響のある気象要因は絶対湿度と言われているが，本研究では温度の相関が高い結果となった．

#### 2.1.3 季節性を反映したモデル

ここでは，予測モデルのデザイン［6］を参考として，短期間の季節性を反映するデータを用いて解析をする．はじめに式(1)で示すモデルを考える．前週の報告数は発表までに遅れがあるため，使用しなかった．

$$[\text{患者数}] = A\times[\text{前々週の報告数}] + B\times[\mathbf{3}\text{週前の報告数}] + C \quad (1)$$

茨城県の 2014 年のデータ(671~700 週)から$A, B, C$を求め，2015 年のデータ(725~750 週)に適用した．さらに上記モデルに新たに気象要素である絶対湿度を加えたモデル式(2)を用いて，上記モデルとの比較を

行った(図 4).

$$[\text{患者数}] = A\times[\text{前々週の報告数}] + B\times[\textbf{3}\text{週前の報告数}] + C\times[\text{前週の気象データ}] + D \quad (2)$$

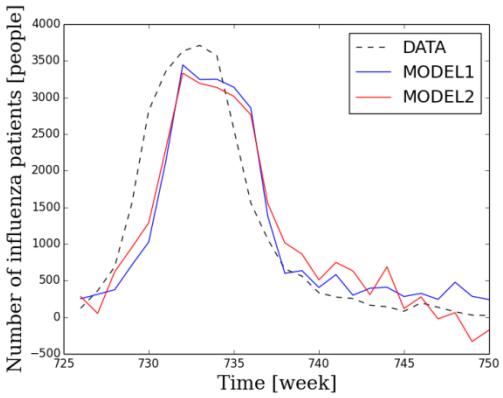


図 4 モデルによる予測結果

気象を考慮することにより，若干だが予測精度は向上した．前年度のデータのみにモデルのパラメータが依存してしまうため，予測の際に前年度のデータの増減の特徴を追従してしまう．そのため，その年特有の大きな変動や 2009 年の新型インフルエンザなどの特性に気をつけなければならない.

### 2.2 機械学習の利用

2.1 にてインフルエンザの季節性を確認し, 気象データを考慮することで，予測精度が向上することを確認した. 本節では平均気温, 最高気温, 最低気温, 降水量の合計，日照時間，平均風速，最大風速，平均蒸気圧，平均湿度といった 9 種類の週ごと気象データを説明変数とし，週ごとの東京都患者数を機械学習によって予測する．今回は，2011 年~2014 年を教師データとして, テストデータとする 2015 年の予測精度を検証する.

#### 2.2.1 主成分分析(PCA)

9 変数からなるデータを，主成分分析によって次元削減を行った．主成分分析とは，多数の変数の相関構造に基づいて少数で全体のばらつきをよく表す主成分と呼ばれる量を合成する多変量解析の一手法である．データに対する説明能力を可能な限り主成分に持たせるために，第一主成分は主成分の分散を最大化するようにして決定する．続く主成分については，それまでに決定した主成分と直交するという制約条件を付け加え，決定する．i 番目の主成分方向への分散の大きさは固有値$\lambda_i$で表される．また，式(3)のように各固有値をすべての固有値の和で割ったものを寄与率と呼ぶ．この寄与率の和で表される累積寄与率は次元の数を決定する基準に用られる．例えば，i 番目までの累積寄与率が 0.9 であれば，i 番目までの主成分で 90%説明していることとなる.

$$\lambda_i / \sum_j^p \lambda_j \quad (3)$$

本研究において，9 変数を主成分分析した結果を表 3，図 5 に示す.

表 3　寄与率と累積寄与率

| | 固有値 | 寄与率 | 累積寄与率 |
|---|---|---|---|
| 1 | 0.350603 | 0.727124 | 0.727124 |
| 2 | 0.062345 | 0.129299 | 0.856423 |
| 3 | 0.036753 | 0.076223 | 0.932647 |
| 4 | 0.010716 | 0.022224 | 0.95487 |
| 5 | 0.008886 | 0.018429 | 0.973299 |
| 6 | 0.006063 | 0.012575 | 0.985874 |
| 7 | 0.00428 | 0.008876 | 0.99475 |
| 8 | 0.002067 | 0.004286 | 0.999036 |
| 9 | 0.000465 | 0.000964 | 1 |

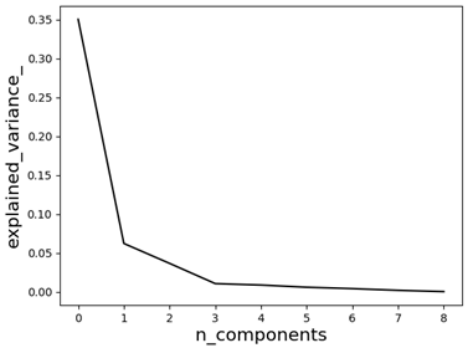


図 5　固有値

3 番目の主成分までで累積寄与率が 0.95 と十分説明出来ているため，3 番目の主成分まで使用することとする.

### 2.2.2 サポートベクター回帰

サポートベクター回帰(以下 SVR)とは分類問題において，サポートベクターマシン（以下 SVM）を回帰問題へ拡張したものである［7］．SVM とは，教師あり学習を用いるパターン認識モデルの一つであり，マージン最大化という基準によってクラス分類を行う超平面を決定する. SVM では高い汎化能力が示されているため，SVR でも高い汎化能力が期待される．

主成分分析により 3 次元に縮約した説明変数において，SVR による予測結果を図 6 に，線形回帰による予測結果を図 7 に示す.

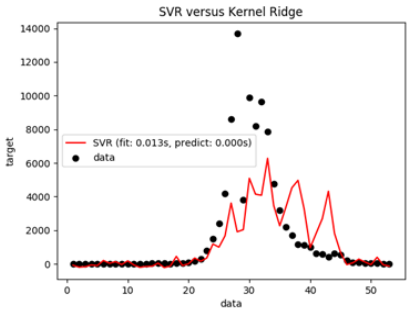


図 6　SVR による予測結果

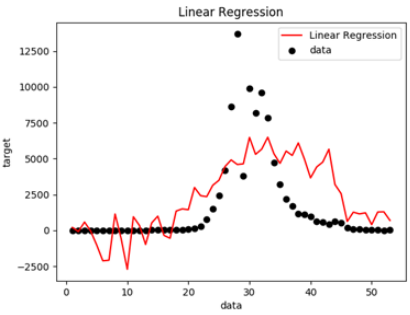


図 7　線形回帰による予測結果

SVR を用いることで，予測精度は向上した.

### 2.3 SIR モデル

2.1，2.2 では気象データを考慮した予測手法について検討した. 本節では, 気象データを考慮せずに, 既存モデルを使用した予測について検討する.

### 2.3.1 SIR モデルについて

伝染病流行の数理モデルとして，ケルマックとマッケンリックは SIR モデルを提案した［8］．このモデルでは S, I, R をそれぞれ感受性人口, 感染人口, 隔離人口として式(4)~(6)のような常微分方程式によって表す.

$$\frac{d}{dt}S(t) = -\beta S(t)I(t) \tag{4}$$

$$\frac{d}{dt}I(t) = \beta S(t)I(t) - \gamma I(t) \tag{5}$$

$$\frac{d}{dt}R(t) = \gamma I(t) \tag{6}$$

ここで，$\beta$ は感染率，$\gamma$ は隔離率である．また，SIR モデルにおいて一人の感染者が再生産する二次感染者の平均数を基本再生産数と言い，N を集団の人口とし，式(7)で表される．R＞1 であれば流行が発生し, インフルエンザの基本再生数は一般的に 2-3 程度とされる.

$$R = \frac{\beta N}{\gamma} \tag{7}$$

### 2.3.2 SIR モデルのパラメータフィッティング

本研究において，報告感染者数のデータに対して感染率 $\beta$ と治癒率 $\gamma$ のパラメータフィッティングを行う．$\beta$ と $\gamma$ はシーズンごと，県ごとの 658 通り算出した．ここでは 1 シーズンを，感染者が最大となった週の前後 15 週の計 31 週とした．フィッティング方法は, 式(4)~ (6)において時間間隔 dt を 1 日とし，$\beta$，$\gamma$ を変化させ報告感染者数との平方二乗誤差の和が最小となる際の値とした．フィッティングの一例を図 8 に示す.

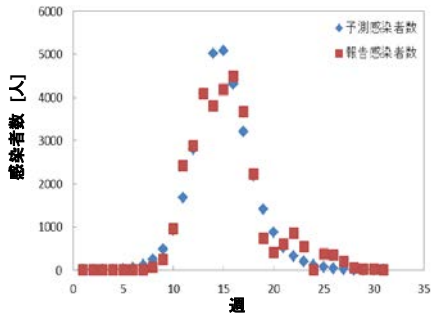


図 8　SIR モデルフィッティングの結果

流行開始時など感染者数が少ない時期にずれがあるが外形は一致していることが確認される．このように算出した $\beta$ と $\gamma$ から，基本再生産数を算出し，報告感染者数との比較を実施した．比較に際して，各県で人口のスケールを揃えるために報告感染者数は各県の人口で割り，感染者率に変換した(図 9).

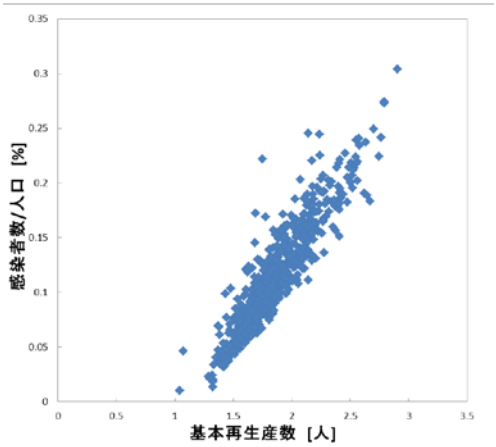


図 9　基本再生産数と感染者率の比較

図 9 において基本再生産数と感染者率の相関係数は 0.92 であり，非常に強い相関が確認される．

### 2.3.3 SIR モデルと感染者予測

2.3.2 では基本再生産数と報告感染者数の相関を示したが，そのシーズンの報告感染者数を既知として SIR モデルへのフィッティングにより基本再生産数を算出しており，感染者の予測を行うには適さない．ここからは流行がピークとなる前の時点までの基本再生産数の値を用いてそのシーズンの総感染者数を予測することを検討する．ここまでの，感染率 $\beta$，治癒率 $\gamma$ は各県，各シーズンで一定としていたが，現実的にはこれらのパラメータは時々刻々と変化していると考えられる．そこで，週ごとに感染率 $\beta$ と治癒率 $\gamma$ を変化させるという条件のもと，再度フィッティングを実施する．また，前処理として，以下の二つのことを行った．

① 各県の人口を 10000 人に揃える．
② 報告感染者数を区間数 3 週として移動平均をとることで平滑化する．

①により，出来るだけ各県のデータを同様に扱えるようにし，②により，出来るだけ感染者の急変動が原因で生じるフィッティング誤差を減少させるようにした．一例を図 10 に示す．

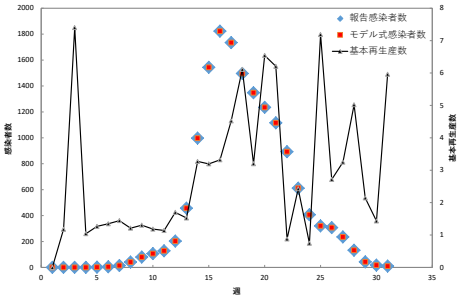


図 10　週ごとにパラメータを変動させた際の結果

全ての週で SIR モデルにより算出した感染者数が報告感染者数に一致していることが確認される．このように算出した基本再生産数が予測に適しているかを調べるため，「ある期間で算出した基本再生産数の平均値」と，「シーズンの総報告感染者数」の 658 プロットについて相関係数を算出する．また，比較対象として「ある期間までの報告感染者数」と「シーズンの総報告感染者数」についても 658 プロットについて相関係数を算出する．「ある期間」とは図 11 に示すように開始週と使用幅を決めた際のデータ区間のことである．まずは一例として開始週を 6 週目，幅 5 週分とした条件で比較を行った(図 12, 図 13).

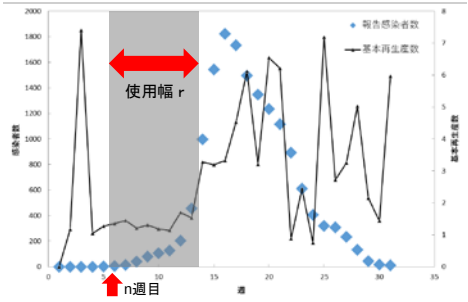


図 11　開始週と使用幅によるある期間の設定

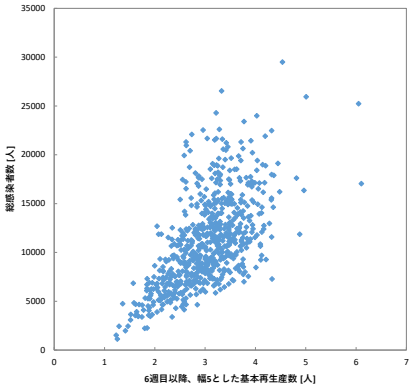


図 12　基本再生産数と総報告感染者数の比較

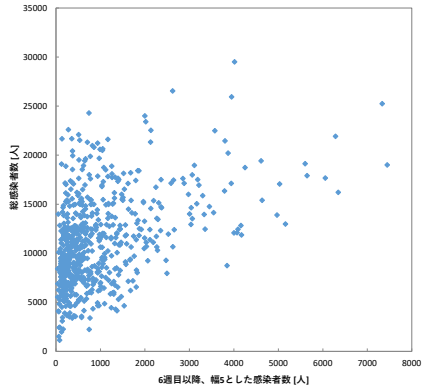


図 13　報告感染者数と総報告感染者数の比較

設定した条件において，基本再生産数の平均値と報告総感染者数の相関係数は 0.61，報告感染者数と総報告感染者数の相関係数は 0.44 となったため，基本再生産数の平均値を用いる方が高精度の予測を行えると言える．しかし，予測に使用するデータ区間によって，それぞれの方法の予測精度に差が出ることが考えられる．その感度を分析するため，開始週を 1~15, 使用幅を 3～15 と変動させた際の相関係数を算出し，等高線として示す(図 14，図 15).

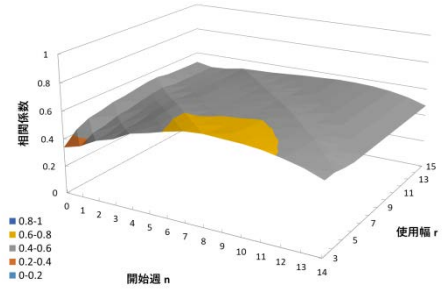


図 14　基本再生産数と総報告感染者数の相関係数

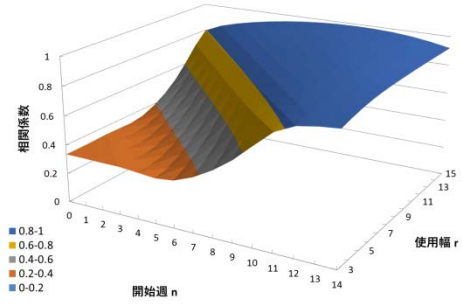


図 15　報告感染者数と総報告感染者数の相関係数

図 14，図 15 より，流行の初期段階のデータだけを用いる場合は，基本再生産数との相関係数の方が高くなるが，流行のピークが近づくと，報告感染者数との相関係数の方が高くなることが確認される．よって，感染者数を予測する際，流行の初期段階においてはモデルフィッティングによる特徴量抽出が有効であると言える．

## 3 結論

スペクトル解析の結果インフルエンザの周期性が判明し，予測には同じく周期性を持つ気象データを考慮することが有効であることを確認した．その後の，9 つの気象データを説明変数とし，4 年分の教師データから予測を行う実験では，線形回帰よりも SVR を用いた方が高精度であることを確認した．更に，SIR モデルでのフィッティングから算出できる基本再生産数を用いることは，各シーズンの総感染者数の早期予測に有効であることを確認した．

## 参考文献

［1］　国立感染症研究所 https://www.niid.go.jp/niid/ja/allarticles/surveillance/2270-idwr/nenpou/6980-idwr-nenpo2015.html, 2017/10/03 確認

［2］　厚生労働省 http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/kekkaku-kansenshou/infulenza/index.html 2017/10/12 確認

［3］　気象庁，過去の気象データ http://www.data.jma.go.jp/obd/stats/etrn/index.php, 2017/10/3 確認

［4］　日野幹雄,「スペクトル解析」, 2010, 朝倉書店

［5］　J.C ミラー，村上正康訳,「統計学の基礎」, 1988, 培風館

［6］　澤井 啓介，坂本 亘,「代数変数によるインフルエンザ流行予測の改良」, 日本計算機統計学会シンポジウム論文集，pp.69-72，2017

［7］　A.J. Smoda and B. Schoelkopf, “A tutorial on support vector regression”, NeuroCOLT2 Technical Report, NC2-TR-1998-030, 1998

［8］　W. O. Kermack and A. G. McKendrick, “A Contribution to the Mathematical Theory of Epidemics, ”*Proc. Roy. Soc. of London. Series A*, Vol. 115, No. 772, pp. 700-721, Aug. 1, 1927