インフルエンザ感染者数の傾向分析と予測

9班

201720612　雨谷　健司　201720622　川﨑　航太

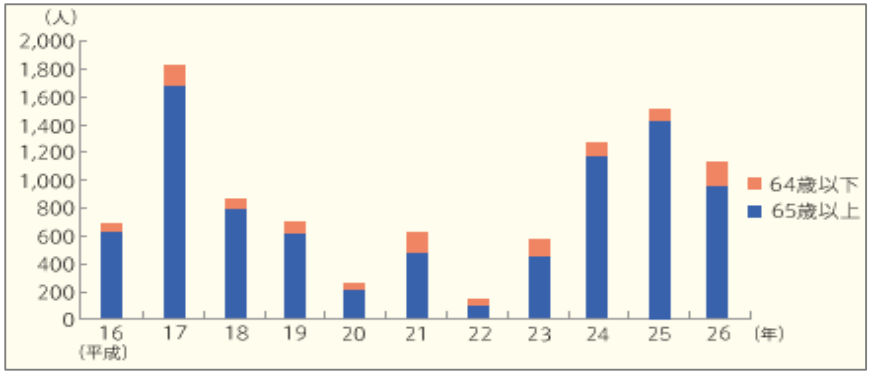
201720633　早瀬　悠希　201720645　楊　明達

指導教官　イリチュ　美佳

筑波大学
University of Tsukuba

研究背景（インフルエンザとは）

原因：インフルエンザウイルス
特徴：ウイルスに型がある
　　感染して免疫を獲得しても何度も感染する
　　合併症・重症化の危険
症状：喉の痛み、咳、鼻水　さらに・・・
　　高熱、全身の倦怠感、頭痛、関節（筋肉）痛
治療法：一般療法と薬物療法
　　←副作用・吸入の必要性・入院の必要性



日本でのインフルエンザによる死亡者数

画像出典 http://www.irasutoya.com/2014/02/blog-post_5079.html
画像出典 https://www.astellas.com/jp/health/healthcare/influenza/basicinformation06.html

研究背景（日本でのインフルエンザ）

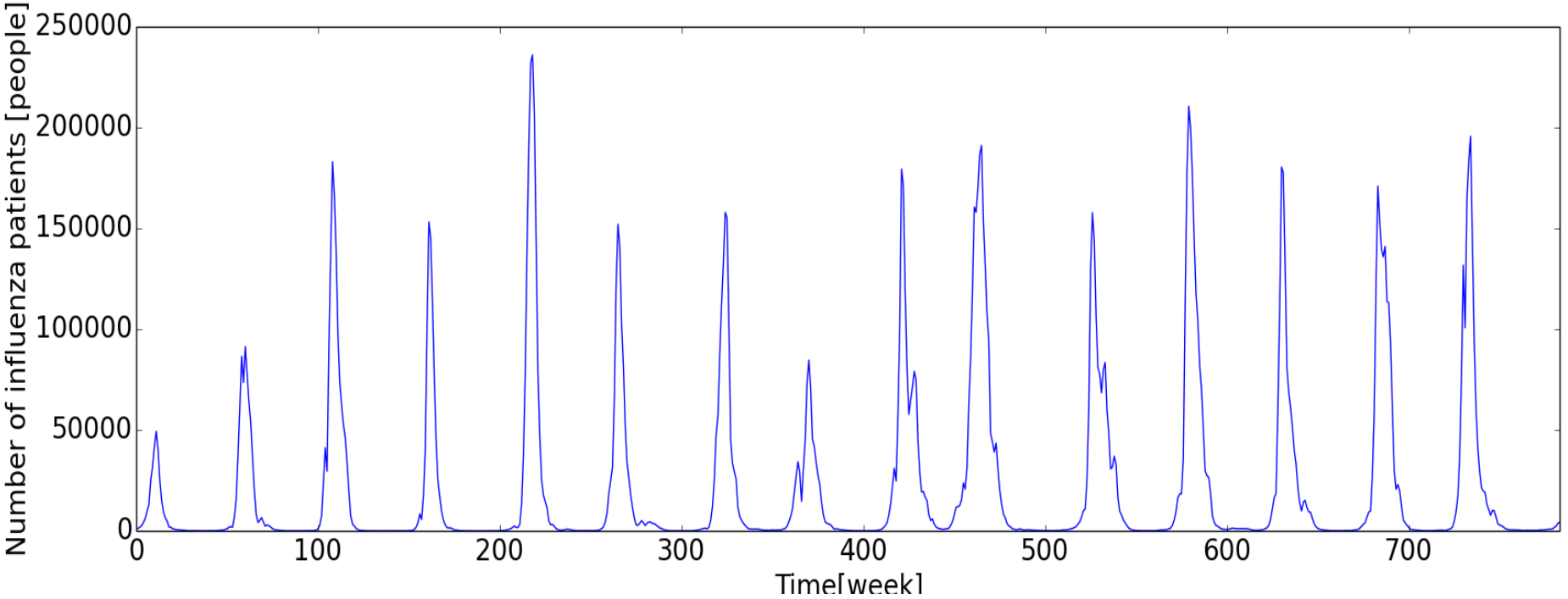
冬(12月から3月)に流行

←インフルエンザウイルスが低温・乾燥を好む

←乾燥した冷たい空気により喉や鼻の粘膜が弱っている

←年末年始の人の移動

画像出展 http://yajibee.com/post-7988/



全国のインフルエンザ患者数の時系列データ

国立感染症研究所のデータを元に作成

研究目的

インフルエンザはとても身近で危険な病気である

目的

流行の予測を行い、今後のリスクを評価

分析・予測手法

1. スペクトル解析 ➡ 周期性の確認
2. 相関分析 ➡ 気象データとの相関確認
3. 季節性反映モデル
4. 主成分分析
5. 機械学習(SVM)

➡ 気象データを用いた予測

6. SIRモデルフィッティング
7. SIRモデルを用いた予測

➡ 気象データを用いず、既存モデルを用いた予測

使用データについて

- 国立感染症研究所が公表しているデータを使用
  (2001年～2015年まで同様の形式で集計されている)
- 現在は「**第10-1表　報告数，週・都道府県・週報定点把握対象疾患・性別**」で検討

2001年　第10-1表　(一部抜粋)

	総　数(total No.)	1週(1week)	2週(2week)	3週(3week)	4週(4week)
	報告数(No. of cases)	報告数(No. of cases)	報告数(No. of cases)	報告数(No. of cases)	報告数(No. of cases)
総　数(total No.)	305,441	1,163	1,875	2,641	4,220
北海道(Hokkaido)	9,635	47	53	61	95
青森県(Aomori)	3,759	10	11	20	23
岩手県(Iwate)	6,072	21	21	23	24
宮城県(Miyagi)	8,866	8	9	28	19
秋田県(Akita)	5,123	13	17	9	23
山形県(Yamagata)	4,178	5	7	10	20
福島県(Fukushima)	5,998	5	10	11	10
茨城県(Ibaraki)	3,642	26	54	67	74
栃木県(Tochigi)	2,337	16	18	22	38
群馬県(Gunma)	4,859	29	27	27	65
埼玉県(Saitama)	19,442	78	136	183	298
千葉県(Chiba)	10,167	38	79	92	168
東京都(Tokyo)	6,114	21	51	73	130
神奈川県(Kanagawa)	15,335	79	120	206	329

・・・

・
・
・

47都道府県×52週(1年)×15年分のデータを使用

スペクトル解析

時系列データ　　スペクトル

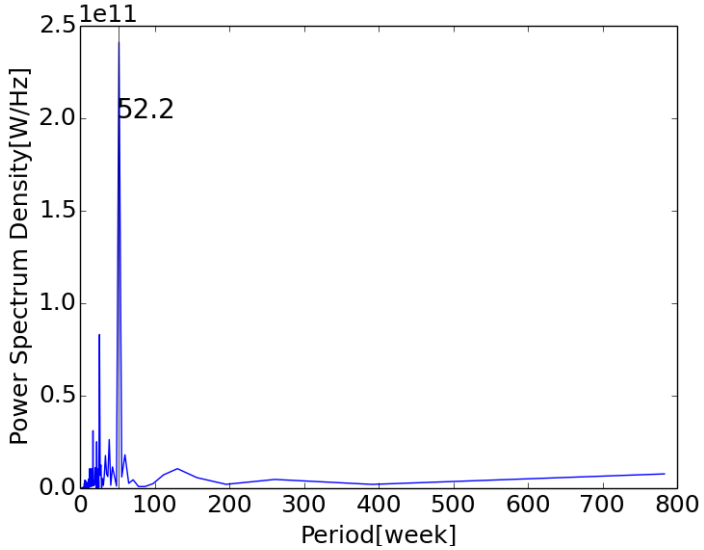
時間領域
(Time domain)

フーリエ変換

フーリエ逆変換

周波数領域
(Frequency domain)

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



全国のインフルエンザの患者数の時系列データのスペクトル解析結果

周期は1年（52.2週）

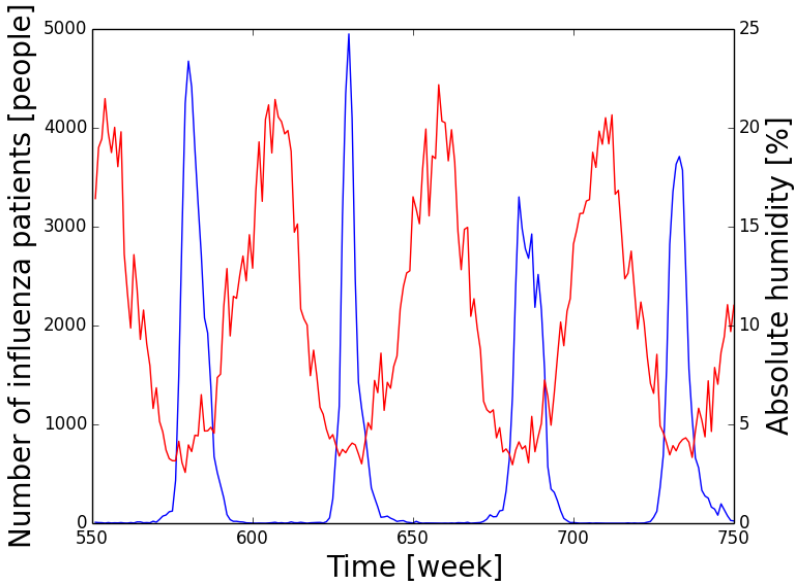
季節性がある

気象データとの関係性

相関分析

2つ以上の変量の間で、一方が変化すると、他方もそれに応じて変化する関係（相関関係）を統計分析すること

相関係数とは、−1から1までの値を取り、絶対値が1に近いほど相関が高い



茨城県のインフルエンザ患者数と絶対湿度の時系列データ

茨城県のインフルエンザ患者数と種々のデータとの時系列と流行期(2015)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
前週の患者数	0.941	0.935
前々週の患者数	0.805	0.769
3週前の患者数	0.639	0.542

気象データでは、温度と絶対湿度の相関が高い
過去のデータとの相関も高い

▽

短期予測へ

気象を考慮した短期予測モデル

MODEL1

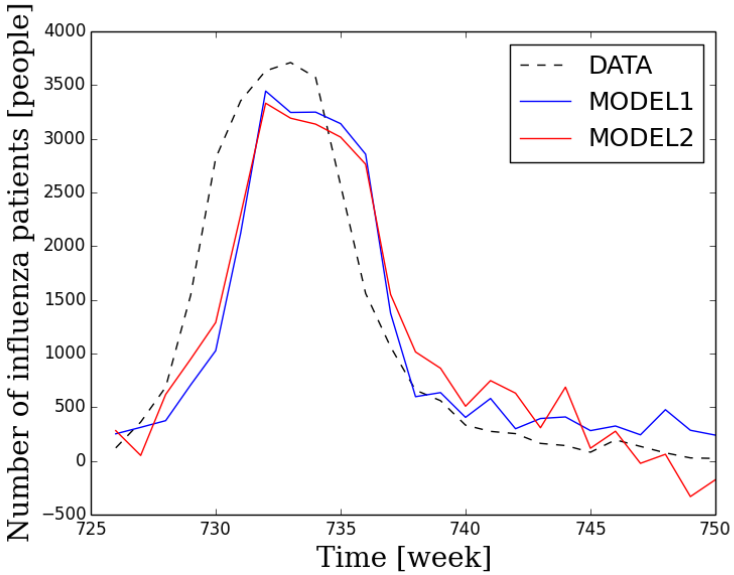
$$[\text{患者数}] = \mathrm{A} \times [\text{前々週の報告数}] + B \times [\text{3週前の報告数}] + C$$

MODEL2

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

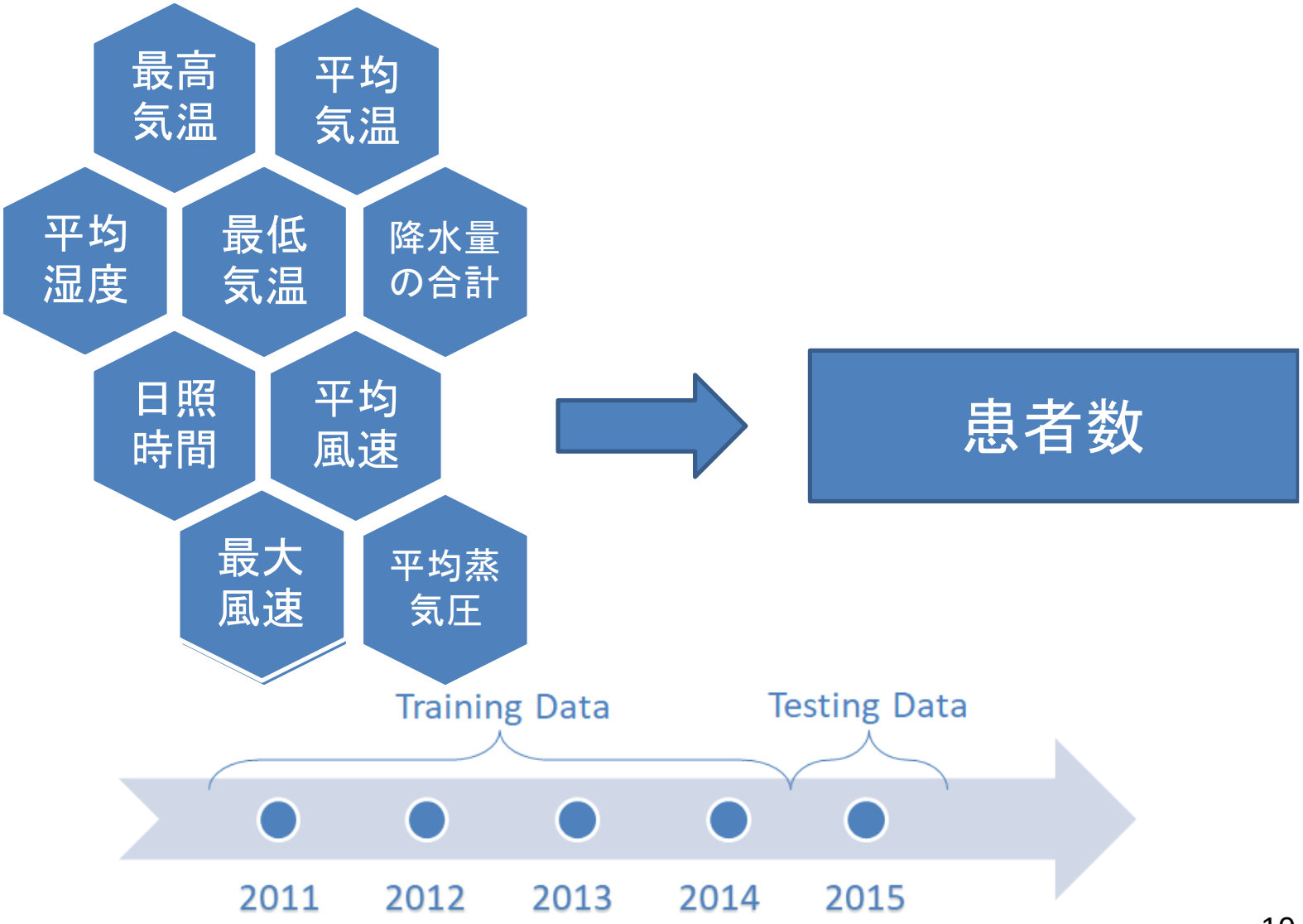
1.前シーズンのデータより、
モデルの係数(A,B,C,D)を決定

2.モデルの入力に今シーズンの
データを使い、予測値を算出



モデルによる茨城県のインフルエンザ患者数の予測結果

機械学習の利用


最高気温
平均気温
平均湿度
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
患者数
Training Data
Testing Data
2011
2012
2013
2014
2015

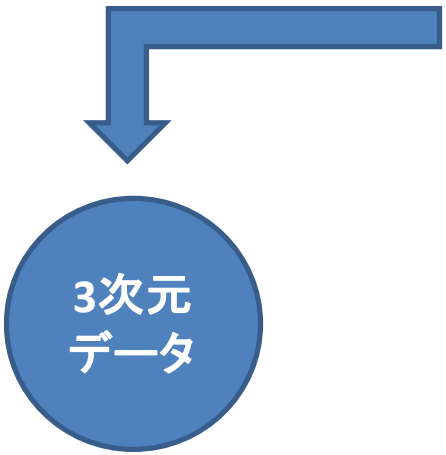
主成分分析(PCA)

主成分分析とは
相関のある多数の変数から相関のない少数で全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する、**データの次元削減**に利用する

固有値
各主成分の分散に対応し、主成分が保持している情報の大きさを示す

寄与率
各主成分が持っている情報の大きさを比率で示す

累積寄与率>90%

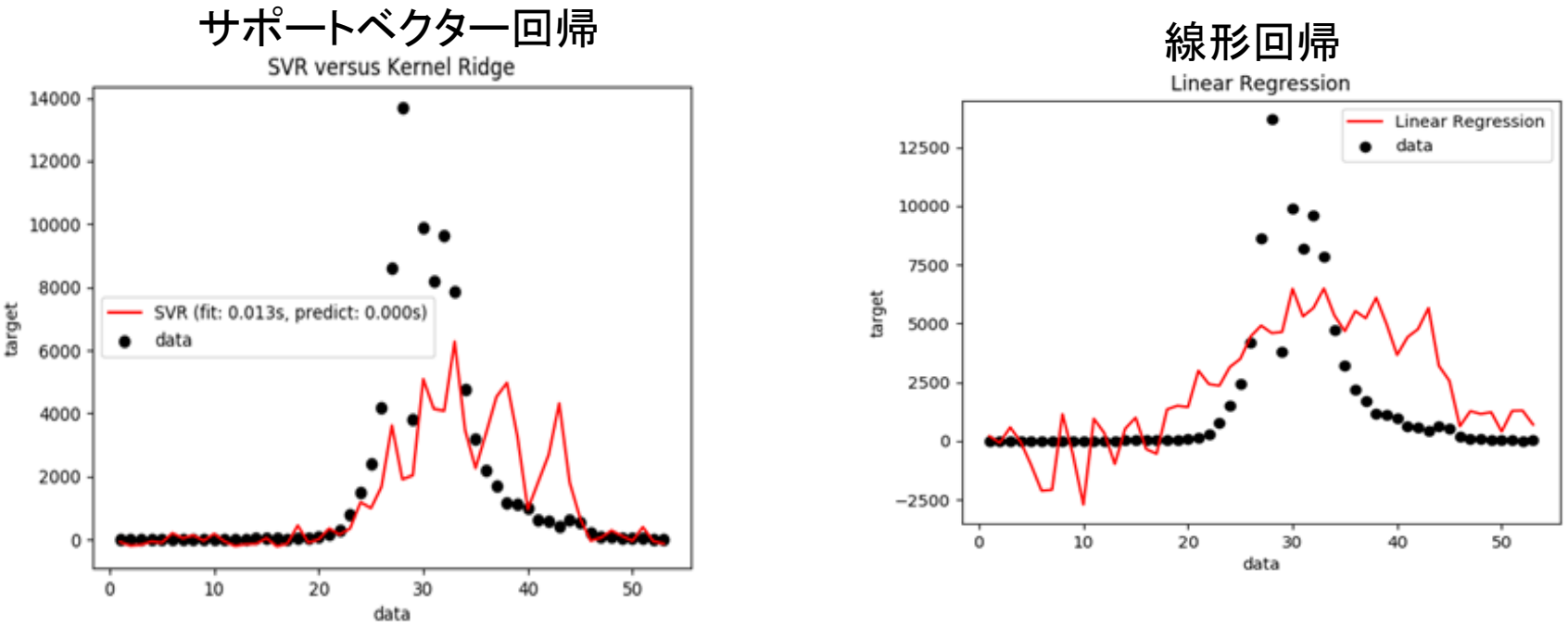


	固有値	寄与率	累積寄与率
1	0.350603	0.727124	0.727124
2	0.062345	0.129299	0.856423
3	0.036753	0.076223	0.932647
4	0.010716	0.022224	0.95487
5	0.008886	0.018429	0.973299
6	0.006063	0.012575	0.985874
7	0.00428	0.008876	0.99475
8	0.002067	0.004286	0.999036
9	0.000465	0.000964	1

サポートベクター回帰(SVR)

SVRとは

サポートベクターマシン（SVM）を回帰問題へ拡張したものである．SVM は，教師付き機械学習を利用した識別器であり，入力となる特徴量の高次元空間における最適な分離超平面を見つけるもので，高い汎化能力が示されている



SVRと線形回帰による回帰分析の比較

SIRモデル

気象データを考慮せず、数理モデルでの予測についても検討

伝染病流行の数理モデルとして、SIRモデルが有名である

(Kermack et.al,1927)

感受性人口:感染可能者
免疫を持たず感染可能(健康な人)
感染人口:感染者
接触した感染可能者に病気を伝染
隔離人口:感染後死亡、もしくは免疫を獲得した人(系から排除された人)

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

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

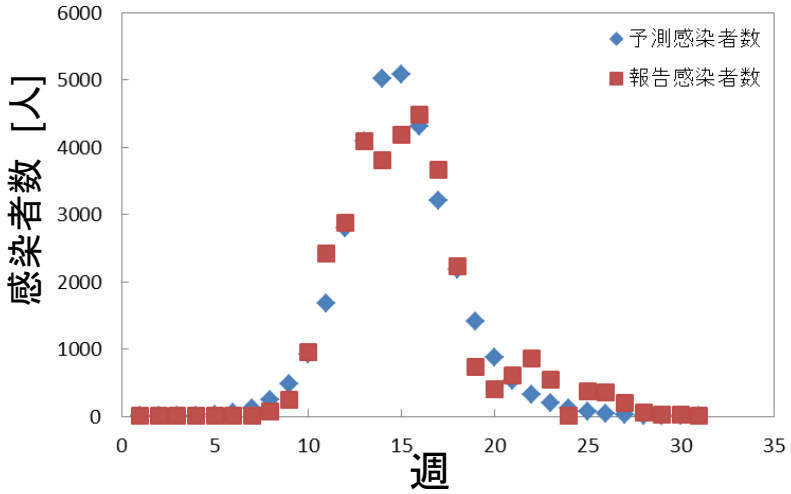
$$\frac{d}{dt}R(t) = \gamma\, I(t)$$

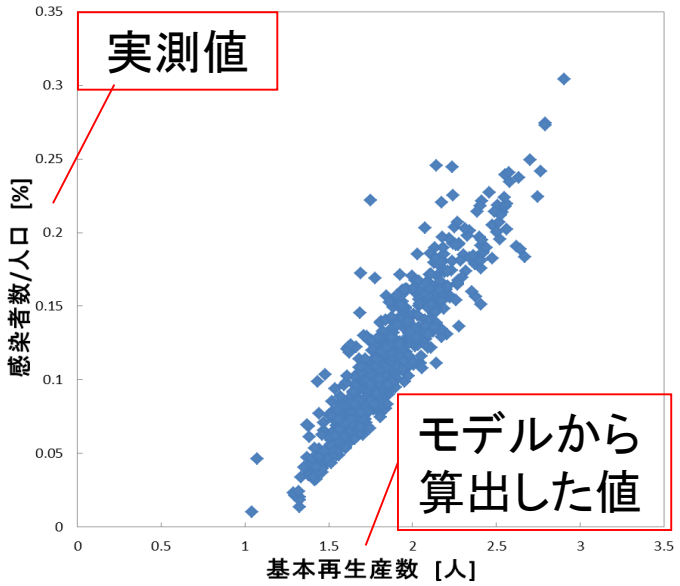
β・・・感染率
γ・・・治癒率

SIRモデル

各県、各年でフィッティングにより、感染率β、治癒率γを決定

例　北海道/2001~2002年 (SIRモデル)





決定した基本再生産数が感染者数と強い相関を持つことを確認

基本再生産数

全体が感受性である人口集団において典型的な1人の感染者が再生産する二次感染者の平均

$$人口 \times \frac{感染率}{治癒率} = 基本再生産数$$

1人の感染者が何人に移すか

SIRモデルを用いた予測

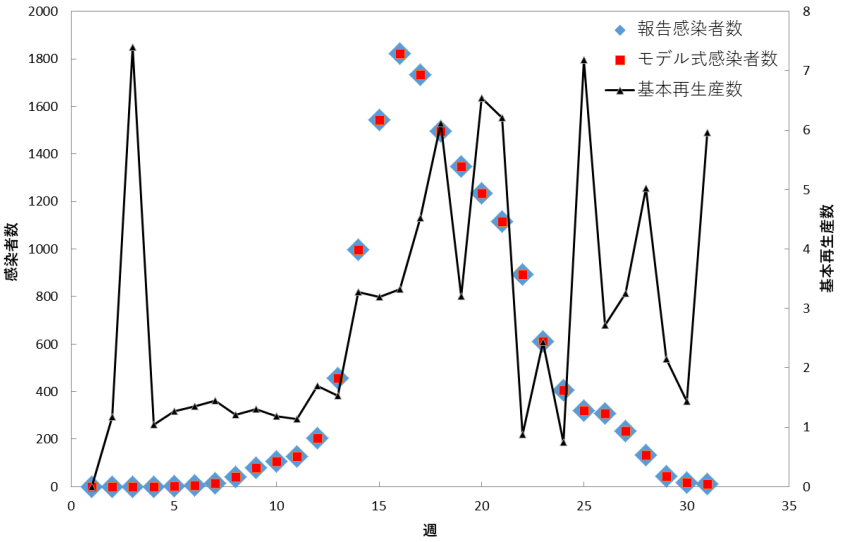
感染者数を既知としてフィッティングしていたため、感染者予測には向かない

感染率β、治癒率γを週ごとに決め、基本再生産数の途中結果を予測に用いる

前処理

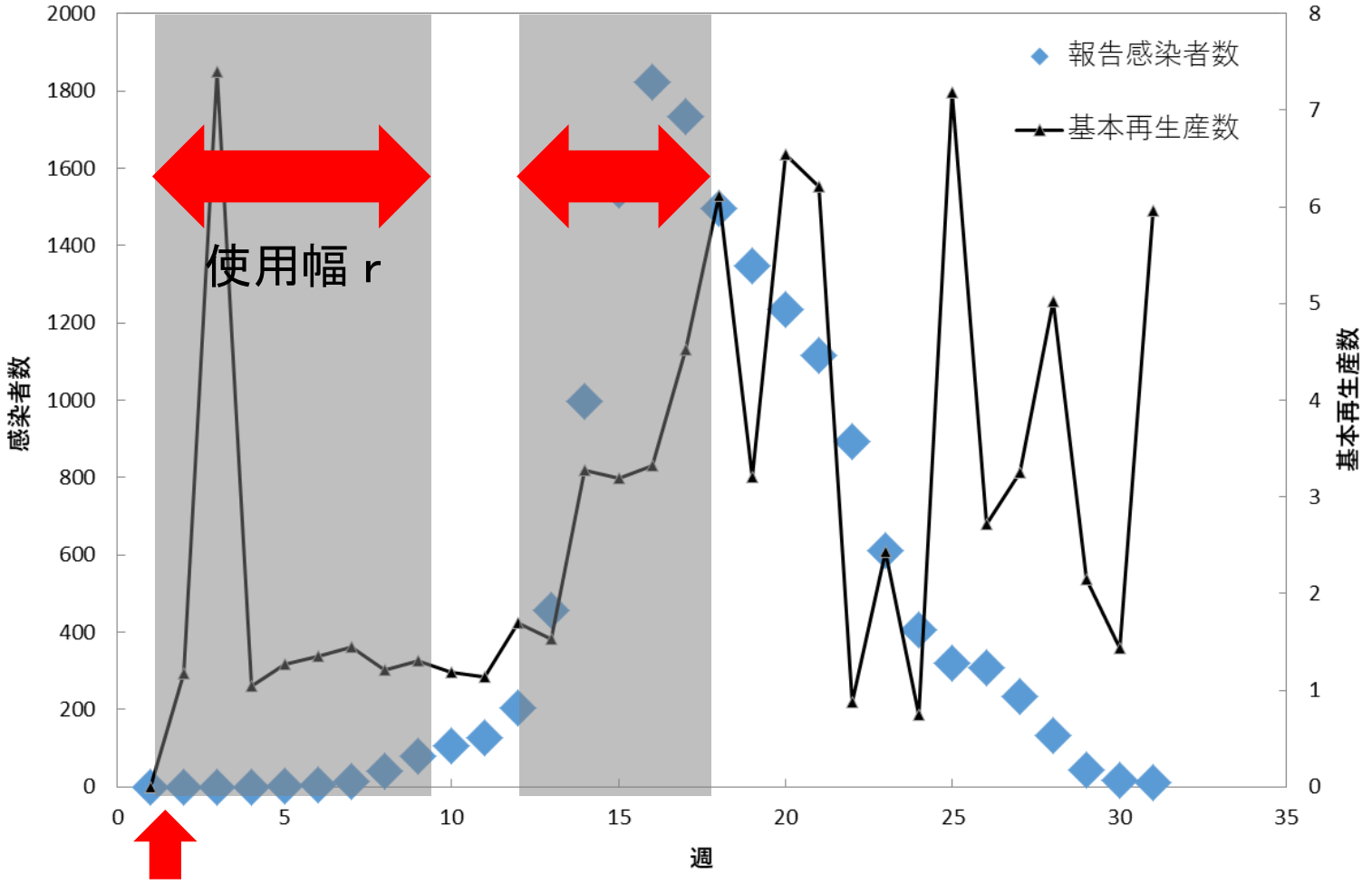
1. 各県の人口を10000人に揃える
   →各県のデータを同等に扱う

2. 報告感染者数を区間数3週として、
   移動平均をとり平滑化する
   →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



週ごとに感染率、治癒率を算出することで、フィッティング精度も向上

感染者予測におけるモデルフィッティングの有効性

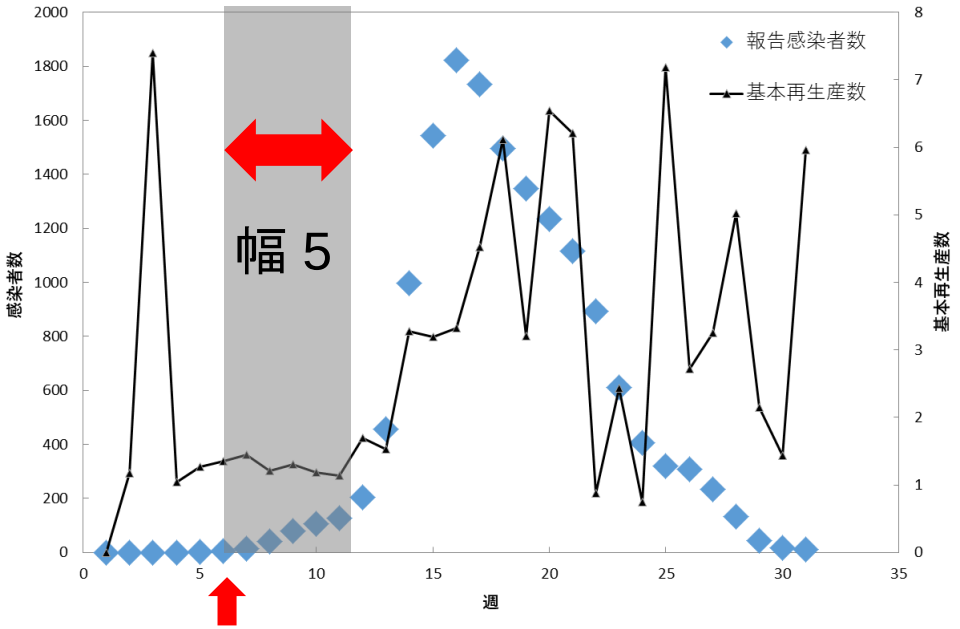


使用データの範囲を変えた際の予測精度を比較・分析

基本再生産数(平均値)　VS　報告感染者数

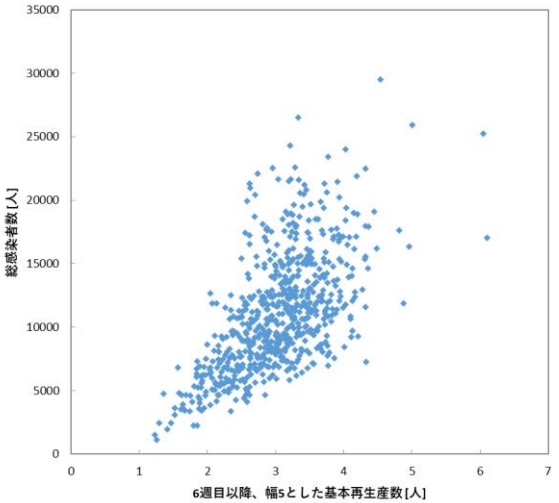
(モデルによる特徴量抽出)　　(元データ)

感染者予測におけるモデルフィッティングの有効性



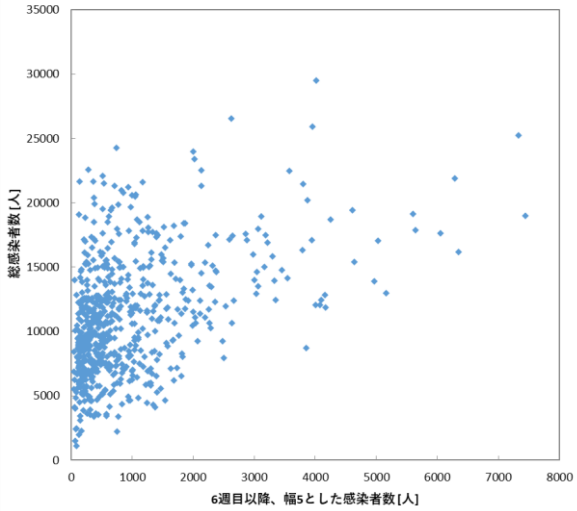
6週目

基本再生産数と報告感染者数



相関係数
0.61

使用期間内の報告感染者数と
総報告感染者数

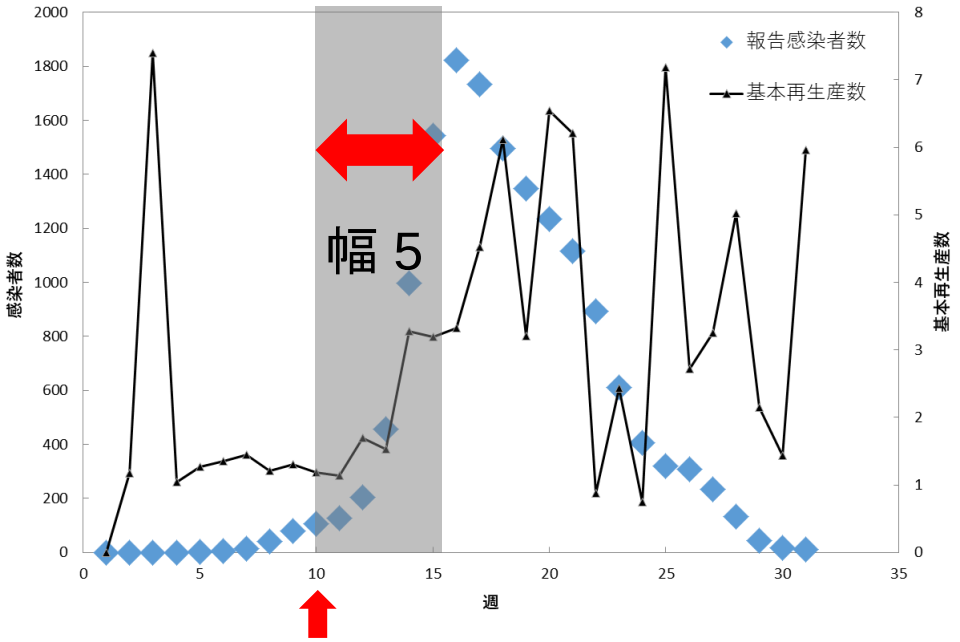


相関係数
0.435

6週目から、使用幅を5と設定した場合

相関
基本再生産数＞報告感染者数

感染者予測におけるモデルフィッティングの有効性



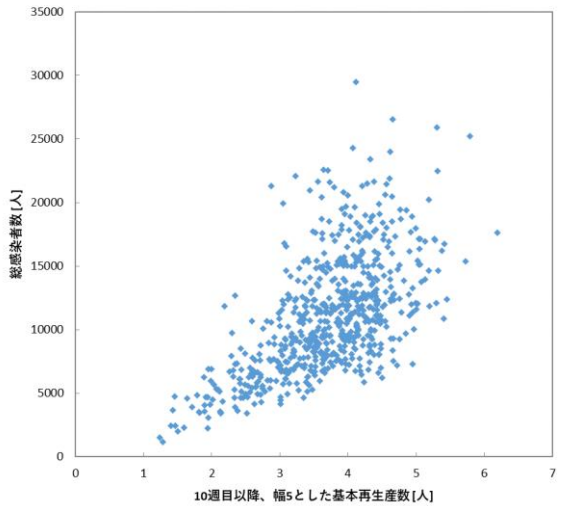
10週目

10週目から、使用幅を5と設定した場合

相関

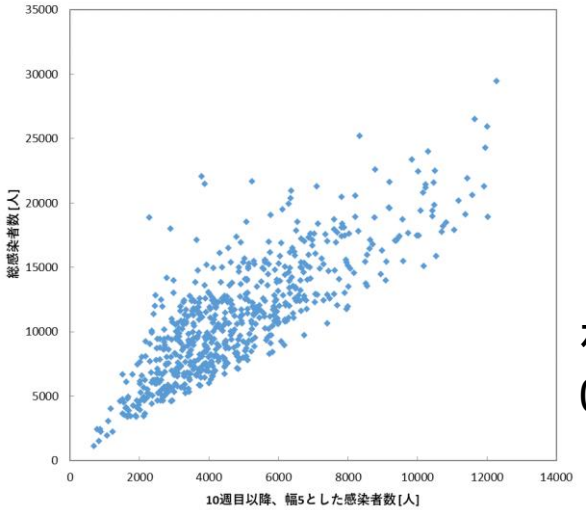
基本再生産数＜報告感染者数

基本再生産数と報告感染者数



相関係数
0.61

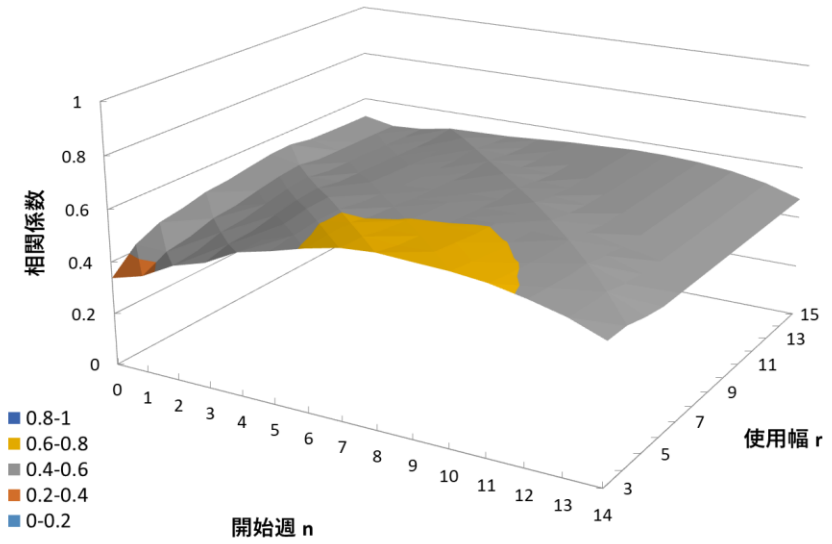
使用期間内の報告感染者数と
総報告感染者数



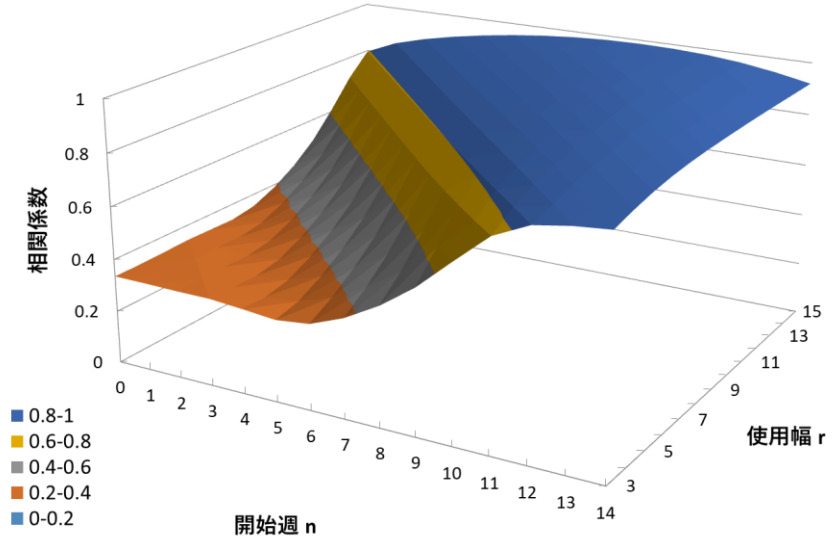
相関係数
0.82

データ使用幅を変えた際の相関

基本再生産数と報告感染者数の相関



使用期間内の報告感染者数と
総報告感染者数の相関



- ✓ 流行の初期段階においては、基本再生産数との相関の方が高い
- ✓ ピークに近づくにつれ、報告感染者数との相関が高くなる

本格的な流行が始まる前のデータからモデルフィッティングにより特徴量を抽出することで、そのシーズンの総感染者数の早期予測が行える

まとめ

- 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た
- 同じく周期性を持つ気象データとの相関が高いことを確認した
- 短期予測モデルでは、気象を考慮することで精度が向上した

- 主成分分析により9つの変数を3つに削減することが出来た
- 線形回帰よりも、SVMを用いた方が結果が向上した

- 気象データを考慮しないSIRモデルは、総感染者の早期予測に有効であった

総まとめ・今後

予測目的に応じた手法の選択が重要
例)来週の感染者を予測したい、今シーズンの総感染者を予測したいなど

気象を用いた予測と数理モデルを用いた予測の融合、
アンサンブル予測なども有効だと考えられる

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. 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
3. 日野幹雄，「スペクトル解析」，2010，朝倉書店
4. J.Cミラー，村上正康訳，「統計学の基礎」，1988，培風館
5. 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
6. A.J. Smoda and B. Schoelkopf, “A tutorial on support vector regression”, NeuroCOLT2 Technical Report, NC2-TR-1998-030, 1998
7. 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 (Aug. 1，1927)，pp. 700-721