

複数時点の単語出現頻度を 扱う時系列データモデリング

磯 颯, 若宮 翔子, 荒牧 英治
奈良先端科学技術大学院大学



SOCIOCOM.jp

NAIST®



Twitter から
インフルエンザ患者数を予測

複数時点の単語出現頻度を 扱う時系列データモデリング

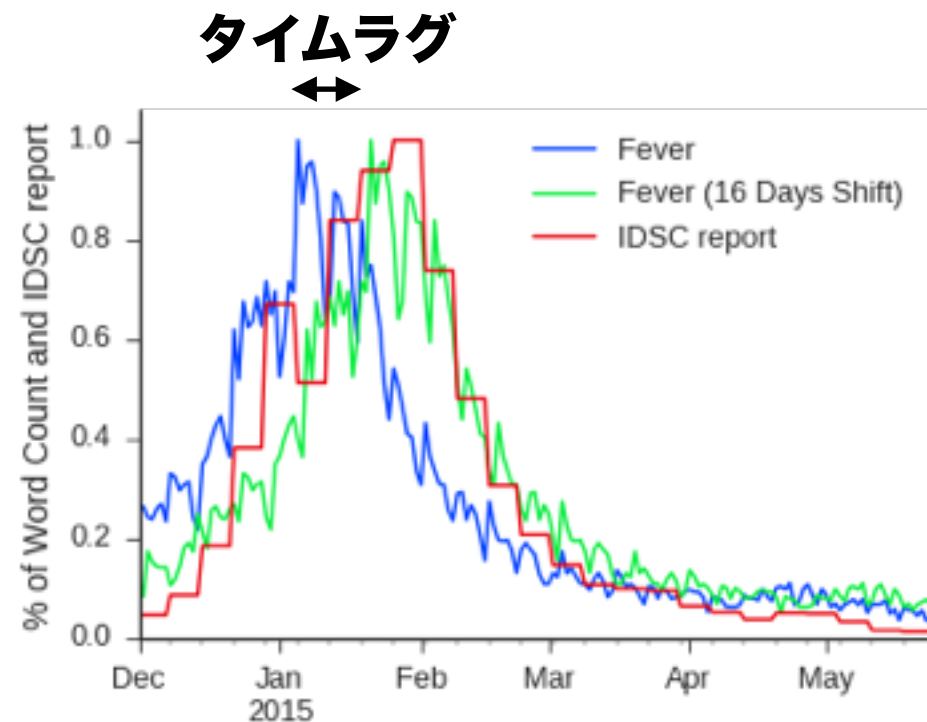
磯 颯, 若宮 翔子, 荒牧 英治
奈良先端科学技術大学院大学



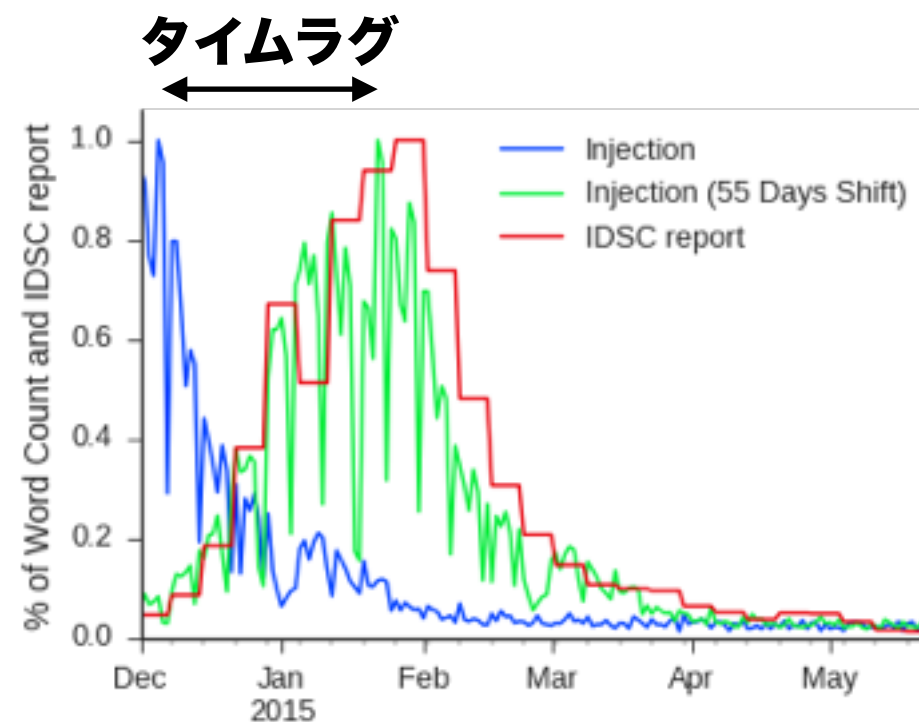
SOCIOCOM.jp

NAIST®

タイムラグのある関係を 自動的に抽出し患者数推定に応用



単語“熱”の発言頻度と患者数



単語“注射”の発言頻度と患者数

■ 発言頻度

■ 提案法

■ 患者数

* 結果概要

1. タイムラグを考慮することで推定精度を改善

2. 予測可能なモデルへ拡張

背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

予測精度

考察

問題点とその改善法

Twitterの利点: 即時性

既存手法:  NIID
NATIONAL INSTITUTE OF
INFECTIOUS DISEASES

国立感染症研究所が、約5,000の協力施設から
週に一度インフルエンザ患者数を集計し、報告

Twitter: 

誰でも利用可能で毎日更新される。

Twitter からインフルエンザの広がりがわかれば、
最新のインフルエンザの患者数を把握できる。

1. 単語頻度の重みづけ和による推定

Lampos and Cristianini, 2010, 他

$$\hat{y}(t) = x_1(t)\hat{\beta}_1 + \dots + x_v(t)\hat{\beta}_v$$

推定患者数 ← 各語の発言回数

2. ノイズ除去 (分類) Aramaki et al., 2011, 他

インフルエンザ関連ツイートを収集し、フィルタリング



普通の人

@ambiguous

人生で一回もインフルに
なったこと無い

削除



インフル患者@ツライ

@flu_patient

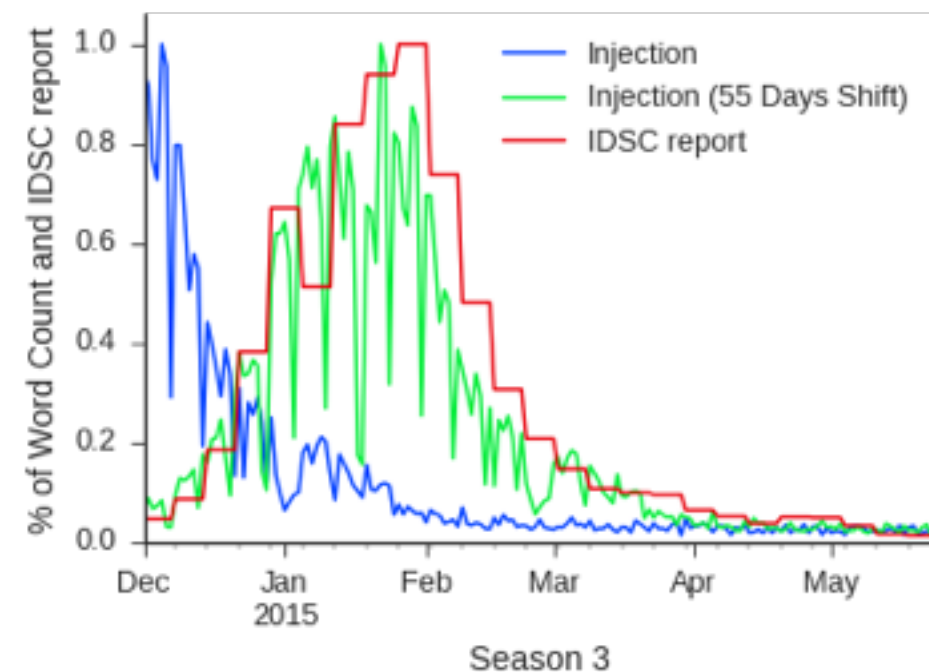
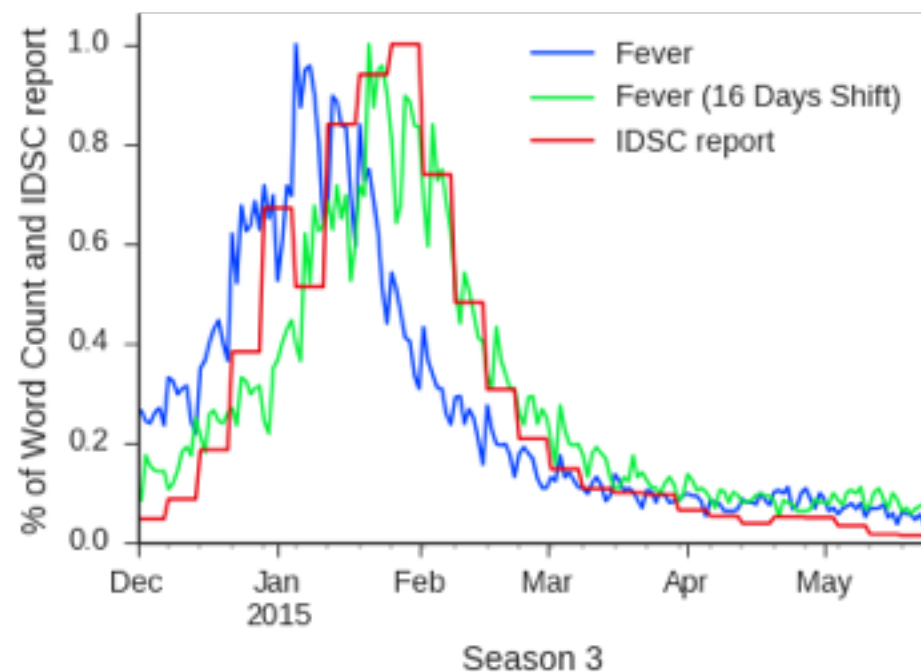
熱めっちゃ上がって来たけど
インフルかなあ

残す

既存手法の問題点

* 仮定:
インフルエンザ関連ツイート $\propto$ 実際の患者数

* 現実:
患者数ピークよりツイート数のピークの方が早い



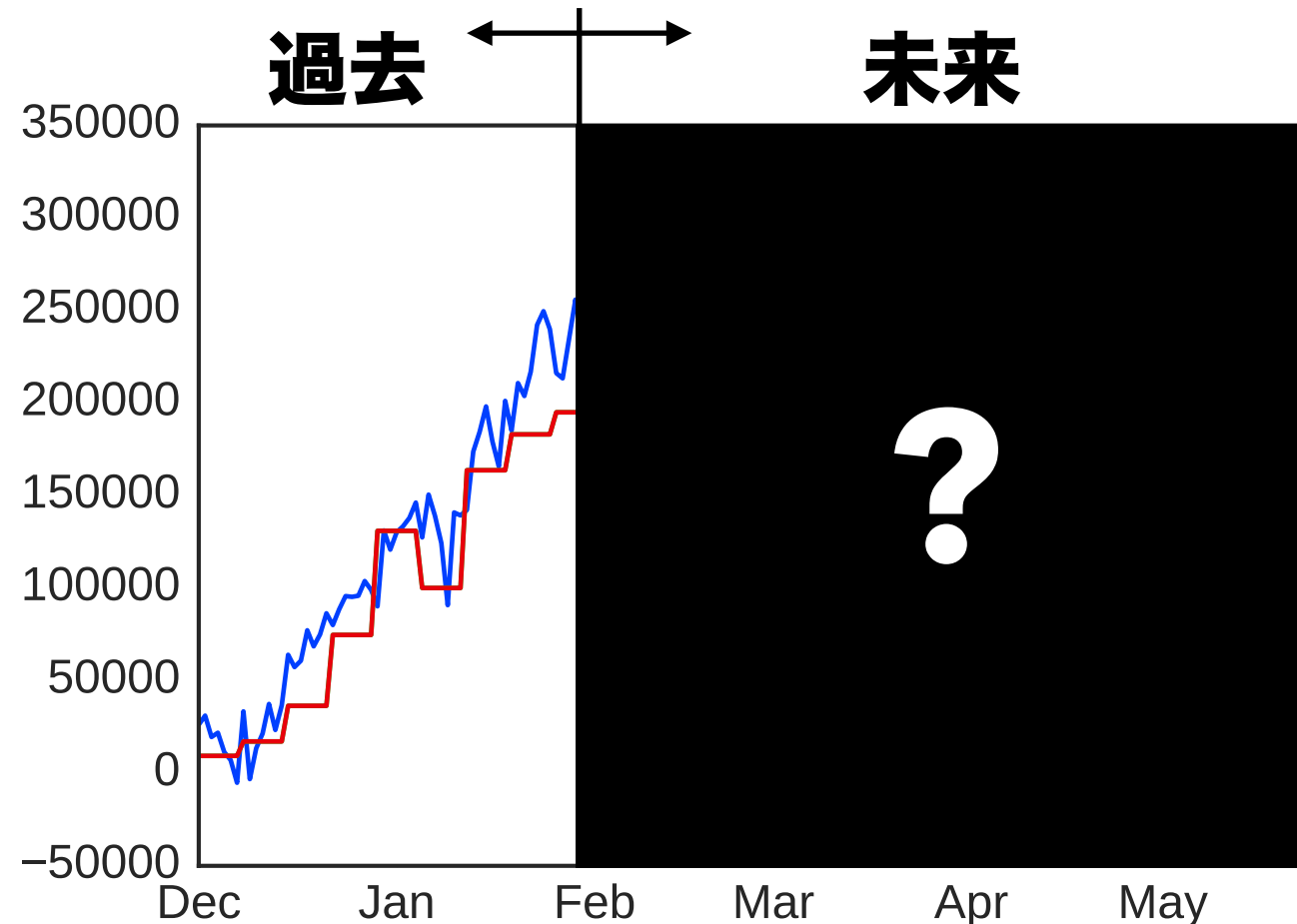
■ 発言頻度, ■ タイムシフト後の頻度, ■ 患者数

既存手法の問題点

予測時点と**同日**の**ツイート**が必要。

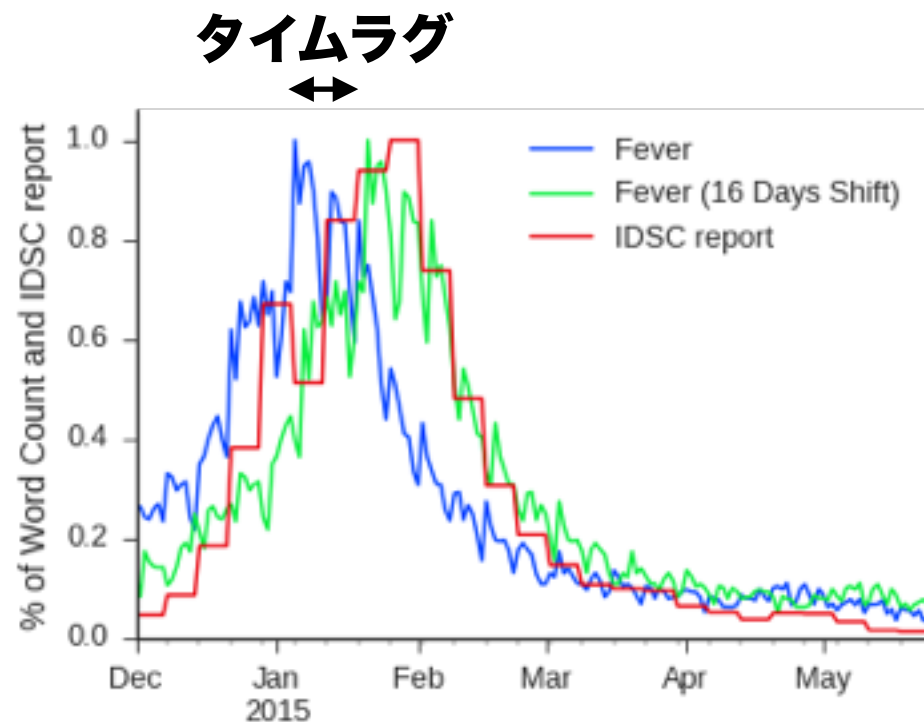


未来の患者数の予測は**出来ない**。

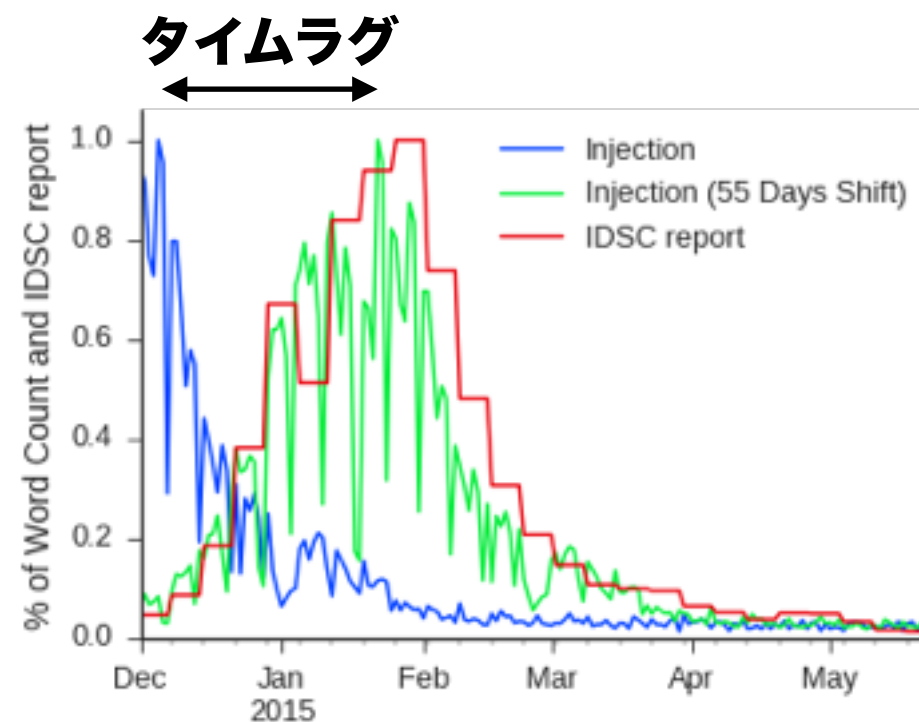


- **予測モデル**
- **患者数**

タイムラグのある関係を 自動的に抽出し患者数推定に応用



単語「熱」の発言頻度



単語「注射」の発言頻度

- 発言頻度
- 提案法
- 患者数

* 結果概要

1. タイムラグを考慮することで推定精度を改善
2. 予測可能なモデルへ拡張

背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

予測精度

考察

問題点とその改善法

タイムラグの指標: 相互相関係数

τ 日前の発言数と**実際の患者数**との
相関関係を測る指標

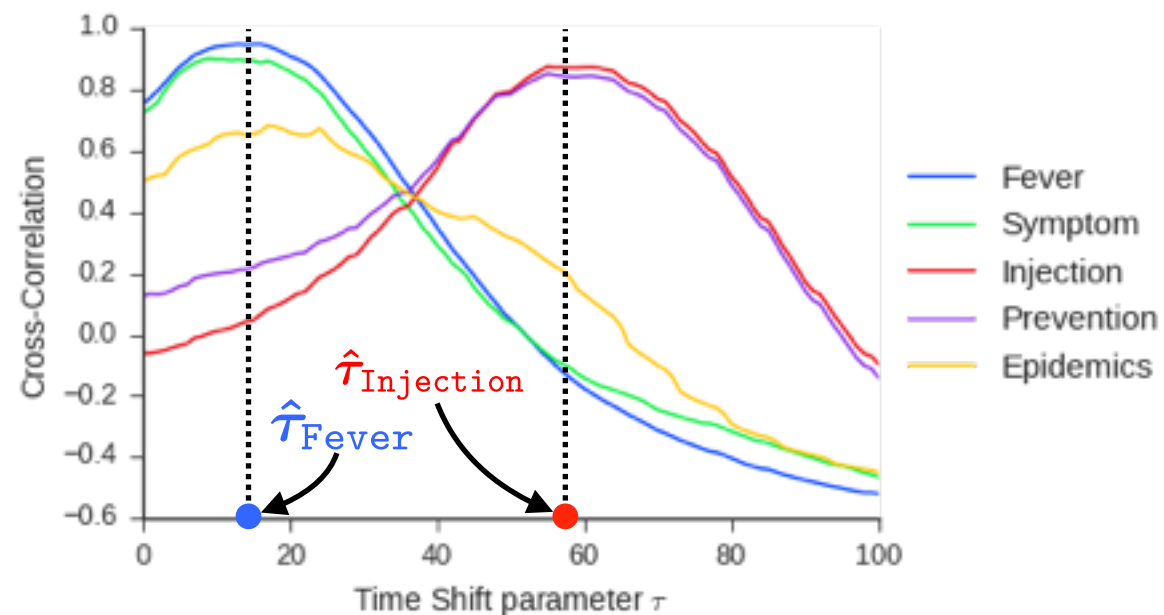
$$r_{x_v, y}(\tau) = \frac{\sum_{t=1}^T (x_v^{(t-\tau)} - \bar{x}_v(\tau))(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^T (x_v^{(t-\tau)} - \bar{x}_v(\tau))^2 \sum_{t=1}^T (y^{(t)} - \bar{y})^2}},$$

ただし $\bar{x}_v(\tau) = \sum_{t=1}^T x_v^{(t-\tau)} / T$

※ $\tau = 0$ ならば, Pearson 相関係数と同値

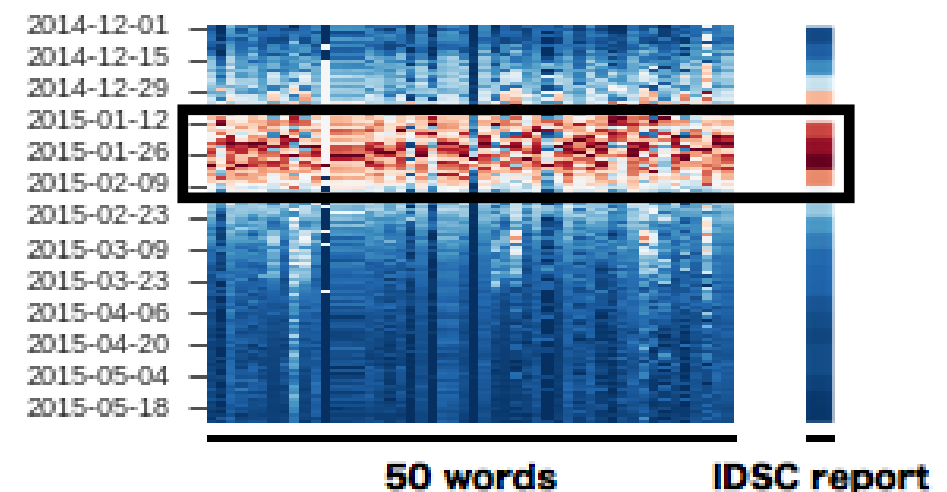
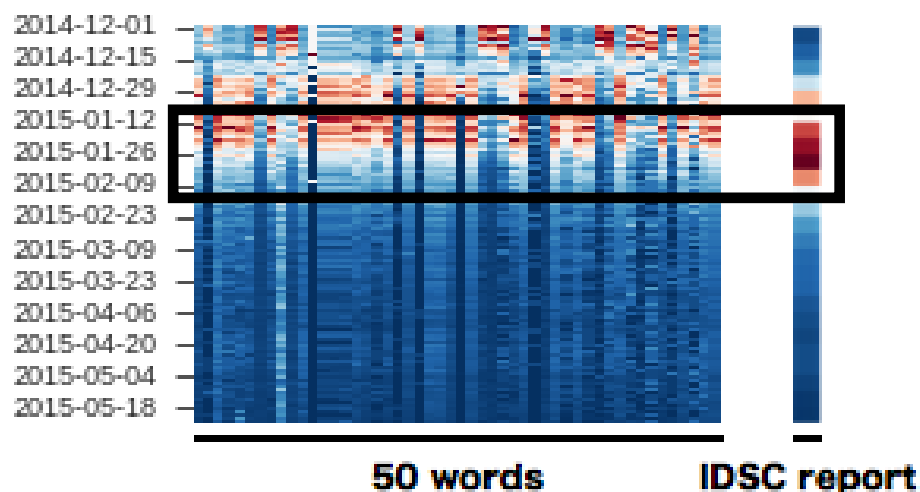
適切なタイムラグ $\hat{\tau}$ を推定

相互相関係数が**最大**となるタイムラグを各語に対して推定



$$\hat{\tau}_v = \arg \max_{\tau} r_{x_v, y}(\tau)$$

タイムシフトにより患者数の傾向を捉えたデータを得る。



1. 単語頻度の重みづけ和による推定

Lampos and Cristianini, 2010, 他

$$\hat{y}(t) = x_1(t)\hat{\beta}_1 + \dots + x_v(t)\hat{\beta}_v$$

推定患者数 ← 各語の発言回数

2. ノイズ除去 (分類)

Aramaki et al., 2011, 他

インフルエンザ関連ツイート 推定には此方を採用



普通の人

@ambiguous

人生で一回もインフルに
なったこと無い



インフル患者@ツライ

@flu_patient

熱めっちゃ上がって来たけど
インフルかなあ

背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

予測精度

考察

問題点とその改善法

* 入力データ: インフルエンザコーパス

- ・ 取得方法: Twitter API から, "**インフル**" を含むツイート
- ・ 期間: 2012年8月 ~ 2016年1月 (3年5ヶ月(現在も収集中))
- ・ ツイート数: 7.7M

前処理

- ・ RT, http を含むツイートを除外
- ・ 形態素解析器: JUMAN (代表表記を単語として用いる.)

* 出力データ: 国立感染症研究所による報告

- ・ インフルエンザの**流行期のみ**患者数を報告している.
- ・ 評価期間を**各年度の流行期**に設定:

Season 1: '12/12/01 ~ '13/05/31, Season 2: '13/12/01 ~ '14/05/31, Season 3: '14/12/01 ~ '15/05/24

* 推定手法

	タイムシフト	
	なし	あり
Lasso (l1) by Tibshirani, 1994	Lasso	Lasso+
Elastic-Net (l1 + l2) by Zou+, 2005	Enet	Enet+

* タイムシフト探索範囲を右記のように設定: $\tau \in \{0, \dots, 60\}$

タイムシフトありで0.111向上

	Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
	Test	Season 1		Season 2		Season 3		
Previous Work	Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.825
	Elastic Net	0.900	0.927	0.809	0.914	0.792	0.805	0.858
Our work	Lasso+	0.952	0.907	0.951	0.888	0.955	0.963	0.936
	Elastic Net+	0.944	0.898	0.960	0.878	0.967	0.959	0.934

※ 推定患者数と実際の患者数との相関係数，高いほど良い



タイムシフトありで0.076向上

	Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
	Test	Season 1		Season 2		Season 3		
Previous Work	Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.825
	Elastic Net	0.900	0.927	0.809	0.914	0.792	0.805	0.858
Our work	Lasso+	0.952	0.907	0.951	0.888	0.955	0.963	0.936
	Elastic Net+	0.944	0.898	0.960	0.878	0.967	0.959	0.934

※ 推定患者数と実際の患者数との相関係数，高いほど良い



タイムシフトなし: Lasso << Enet

	Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
	Test	Season 1		Season 2		Season 3		
Previous Work	Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.825
	Elastic Net	0.900	0.927	0.809	0.914	0.792	0.805	0.858
Our work	Lasso+	0.952	0.907	0.951	0.888	0.955	0.963	0.936
	Elastic Net+	0.944	0.898	0.960	0.878	0.967	0.959	0.934

※ 推定患者数と実際の患者数との相関係数, 高いほど良い



タイムシフトあり：Lasso $\doteq$ Enet

	Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
	Test	Season 1		Season 2		Season 3		
Previous Work	Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.825
	Elastic Net	0.900	0.927	0.809	0.914	0.792	0.805	0.858
Our work	Lasso+	0.952	0.907	0.951	0.888	0.955	0.963	0.936
	Elastic Net+	0.944	0.898	0.960	0.878	0.967	0.959	0.934

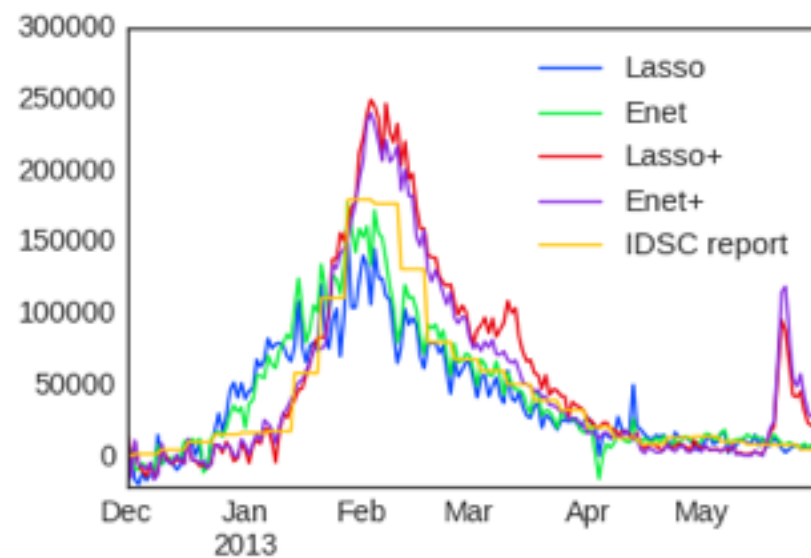
※ 推定患者数と実際の患者数との相関係数，高いほど良い



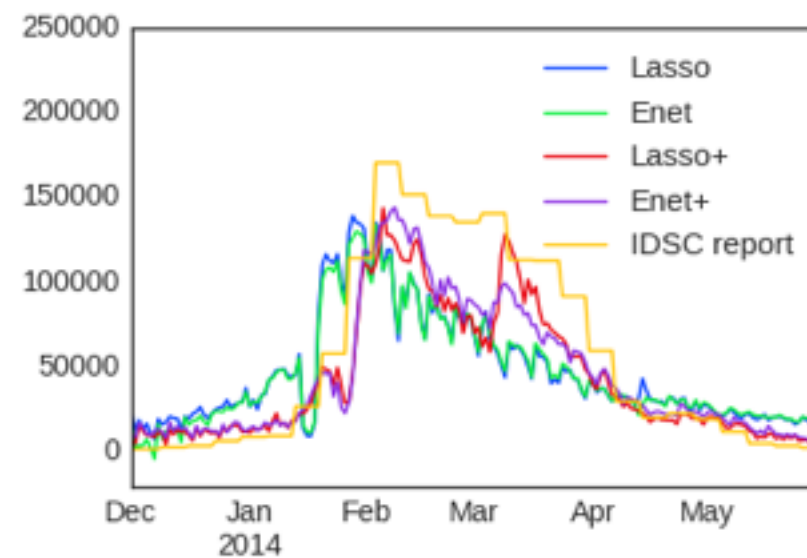
推定結果:

既存法: **Lasso**, **Enet**, 提案法: **Lasso+**, **Enet+**, **患者数**

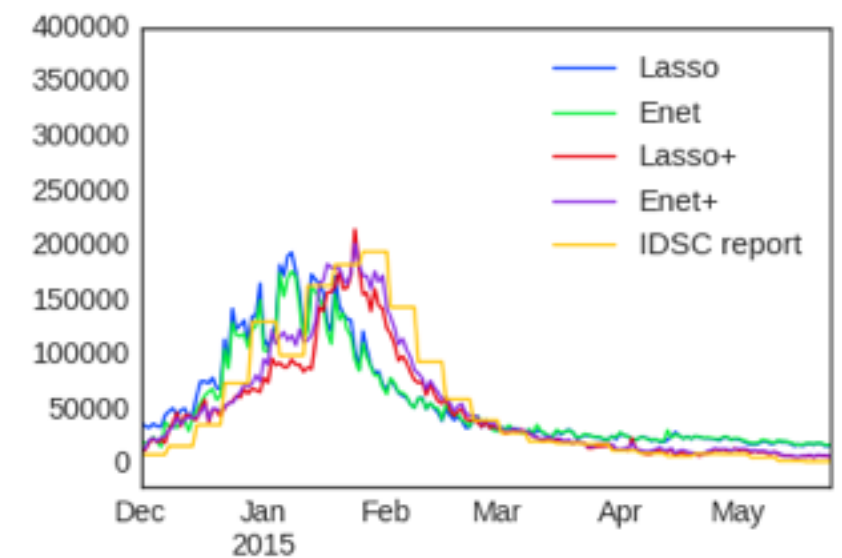
Train: Season 2/ Test: Season 1



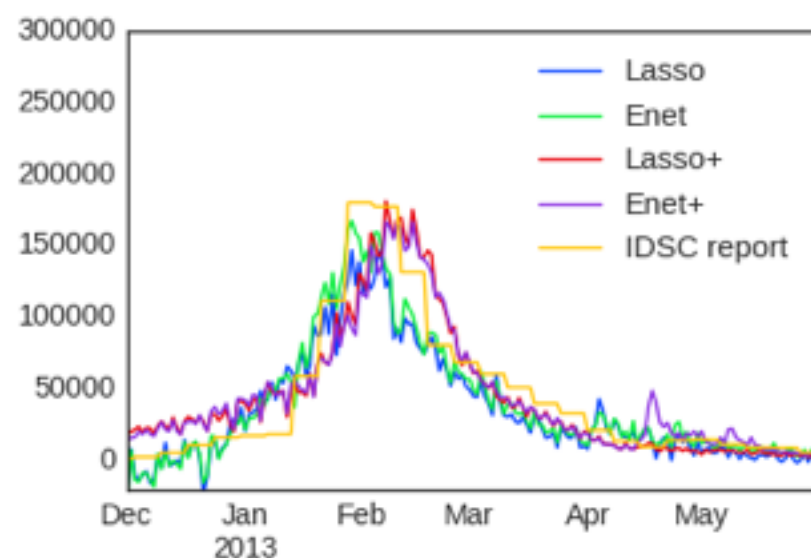
Train: Season 1/ Test: Season 2



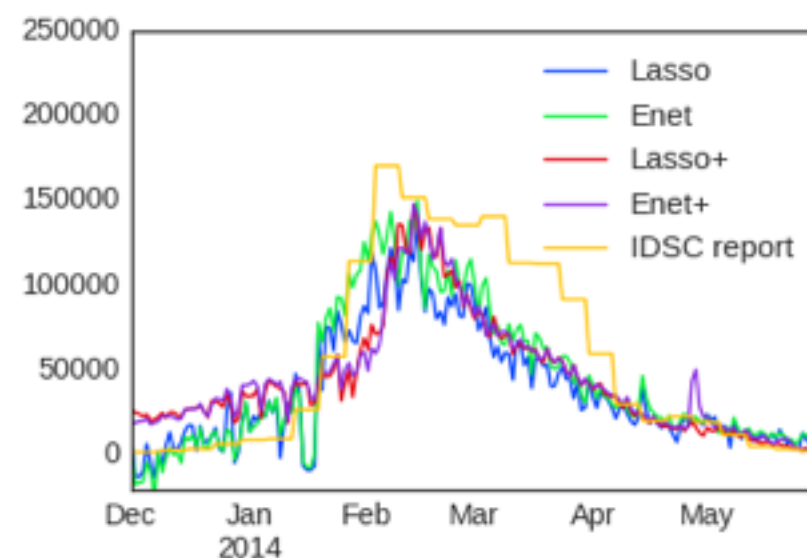
Train: Season 1/ Test: Season 3



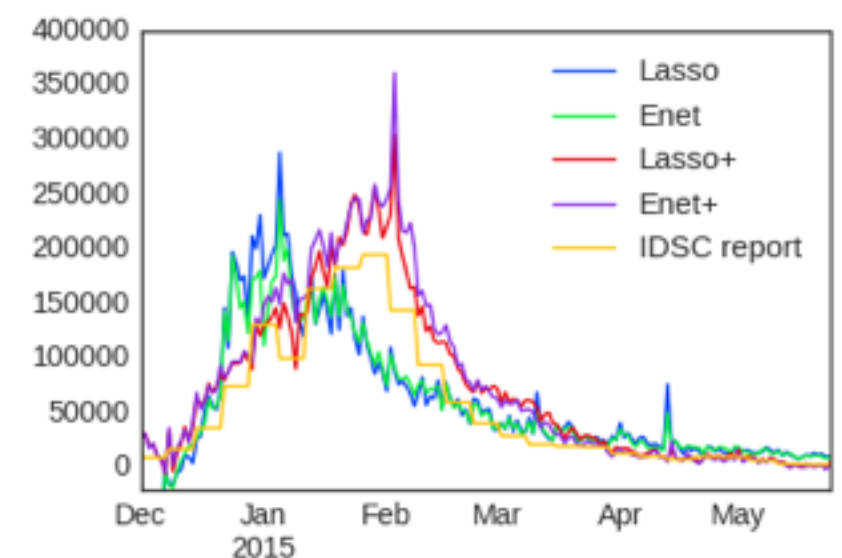
Train: Season 3/ Test: Season 1



Train: Season 3/ Test: Season 2



Train: Season 2/ Test: Season 3



背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

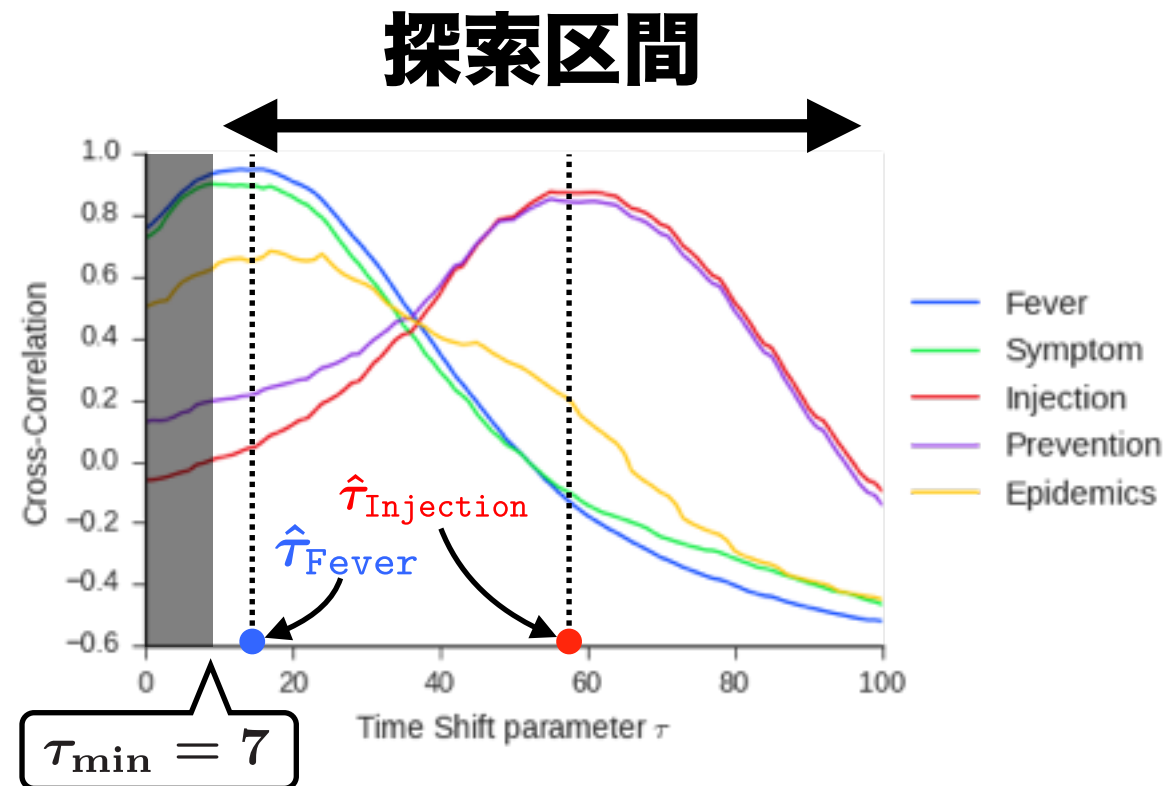
予測精度

考察

問題点とその改善法

e.g. 一週間先の患者数を予測する場合

1. 少なくとも一週間以上のタイムシフトを与える.



$$\hat{\tau}_v = \arg \max_{\tau \in \{7, \dots\} \subset \mathbb{N}} r_{x_v, y}(\tau)$$

2. 観測済みの単語頻度から,

一週間先の患者数の予測モデルを推定出来る.

$$\hat{y}(t+7) = x_1(t-\tau'_1)\hat{\beta}_1 + \dots + x_v(t-\tau'_v)\hat{\beta}_v$$

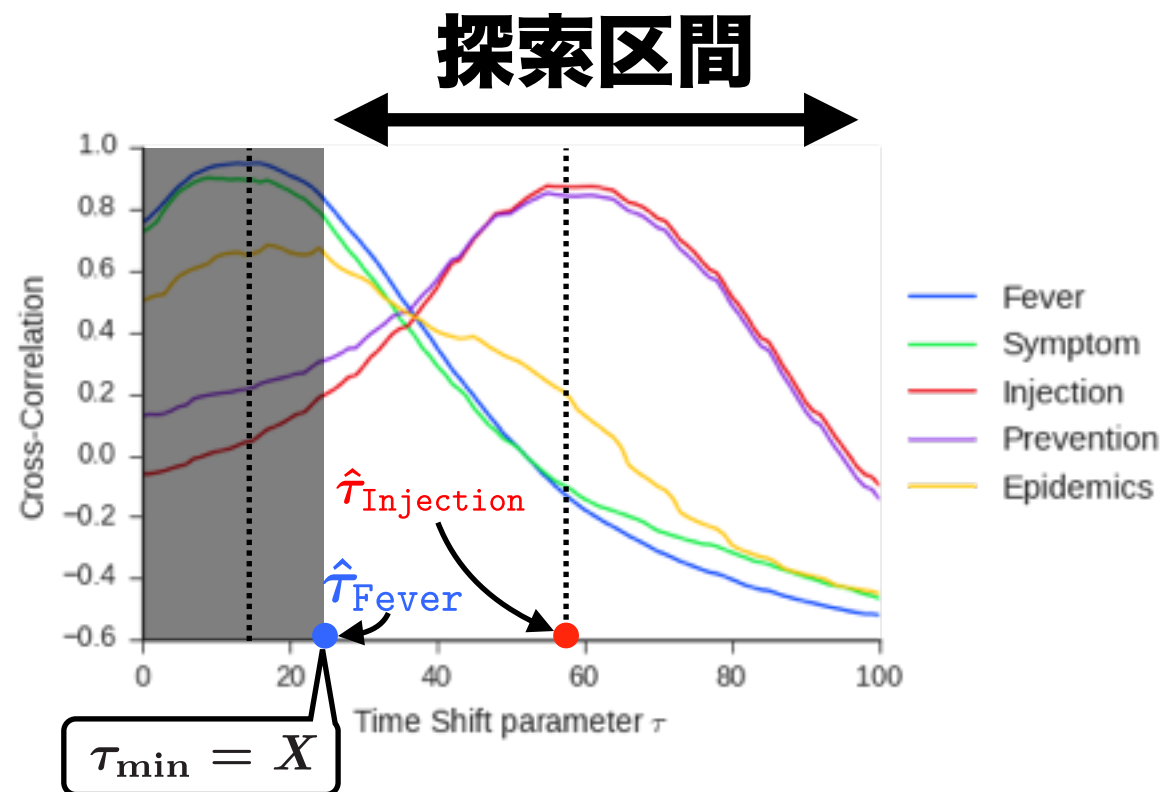
一週間先の患者数

観測済みの単語頻度

※ $\tau'_v = \hat{\tau}_v - 7$

X日先を予測するアルゴリズムに一般化

1. **少なくとも** X日以上 of タイムシフトを与える.



$$\hat{\tau}_v = \arg \max_{\tau \in \{\tau_{\min}, \dots\} \subset \mathbb{N}} r_{x_v, y}(\tau)$$

2. **観測済みの単語頻度**から,

X日先の患者数の予測モデルを推定出来る.

$$\hat{y}(t+X) = x_1(t-\tau'_1)\hat{\beta}_1 + \dots + x_v(t-\tau'_v)\hat{\beta}_v$$

X日先の患者数

観測済みの単語頻度

※ $\tau'_v = \hat{\tau}_v - X$

背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

予測精度

考察

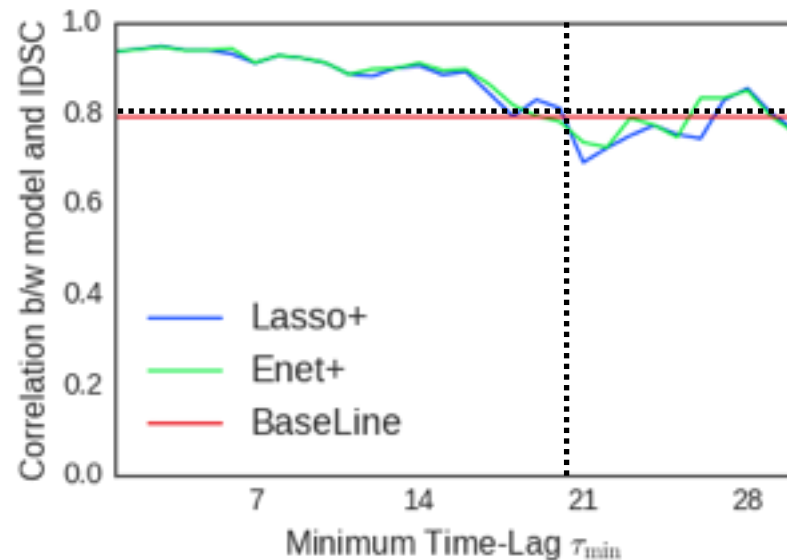
問題点とその改善法

実験設定: 将来の患者数予測

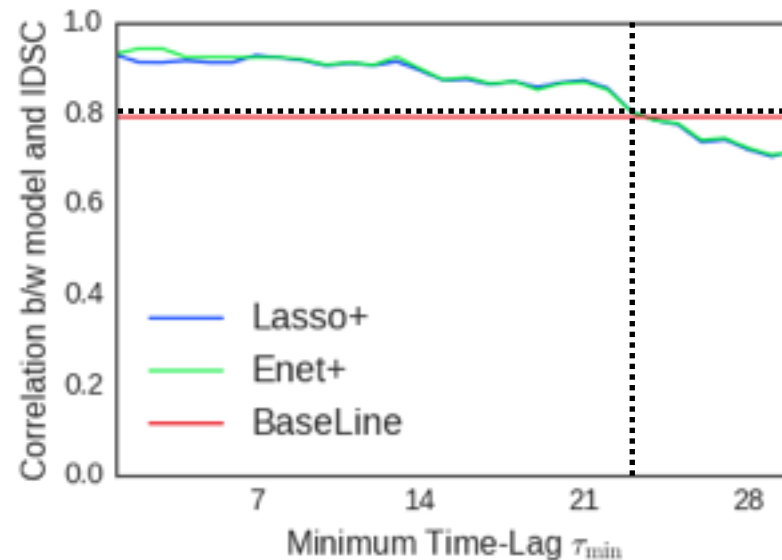
- * 基本的な設定は Experiment 1 と同様
- * 従来法 (Lasso, Enet) は将来の患者数を予測できない
- * 推定手法
 - BaseLine: $\hat{y}_{\text{test}}(t) = y_{\text{train}}(t)$
(学習データ中の実際の患者数をそのまま推定値とする。)
 - Lasso+: タイムシフトありのデータを用いて, Lassoで推定
 - Enet+: 同データをElastic Netで推定.
- * $\tau_{\min}$ を 1 から 30 に変化させ推定精度の推移を検証

Baselineを大きく上回る精度

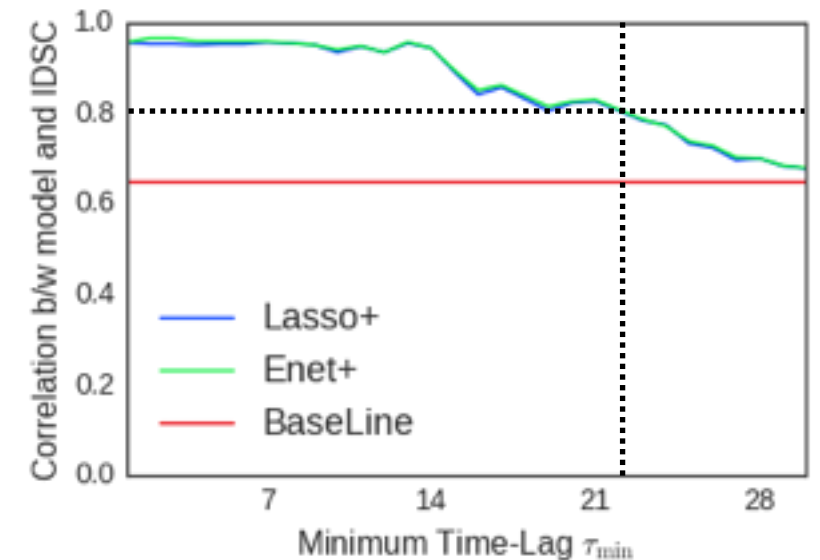
Train: Season 2/ Test: Season 1



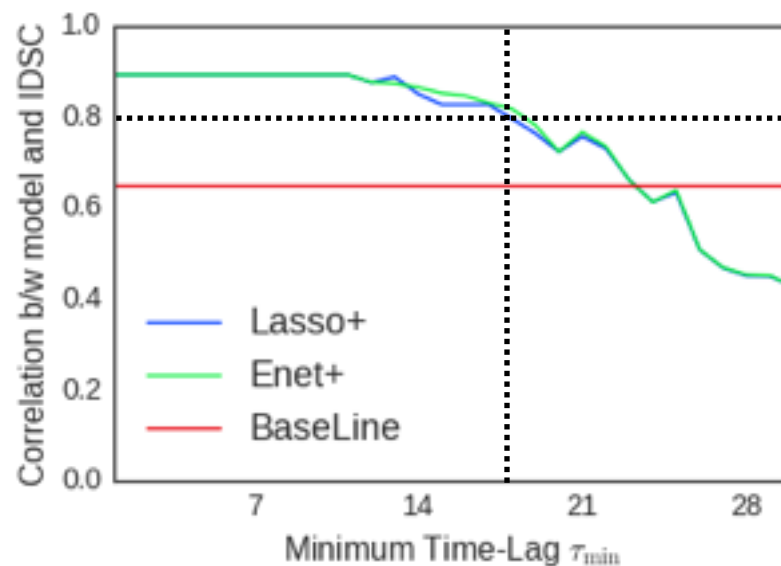
Train: Season 1/ Test: Season 2



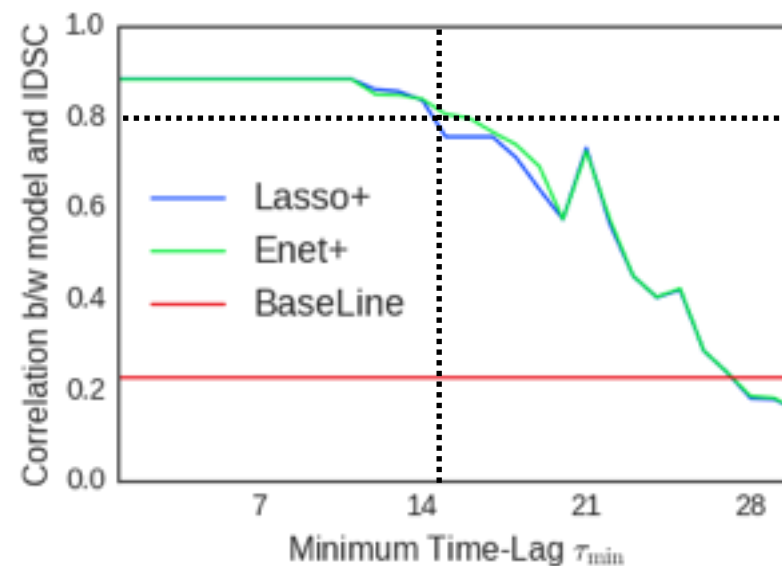
Train: Season 1/ Test: Season 3



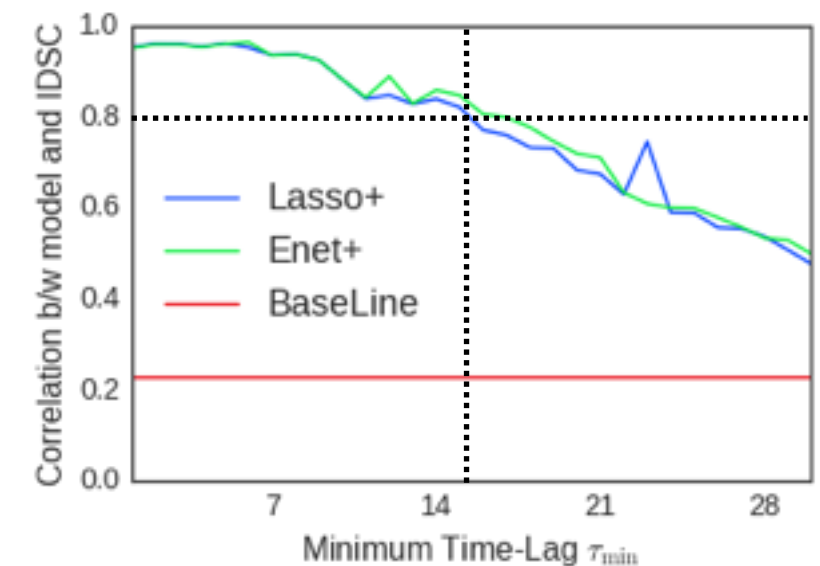
Train: Season 3/ Test: Season 1



Train: Season 3/ Test: Season 2



Train: Season 2/ Test: Season 3



※ 横軸: 予測日数, 縦軸: 推定患者数と実際の患者数との相関係数, 高いほど良い



背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

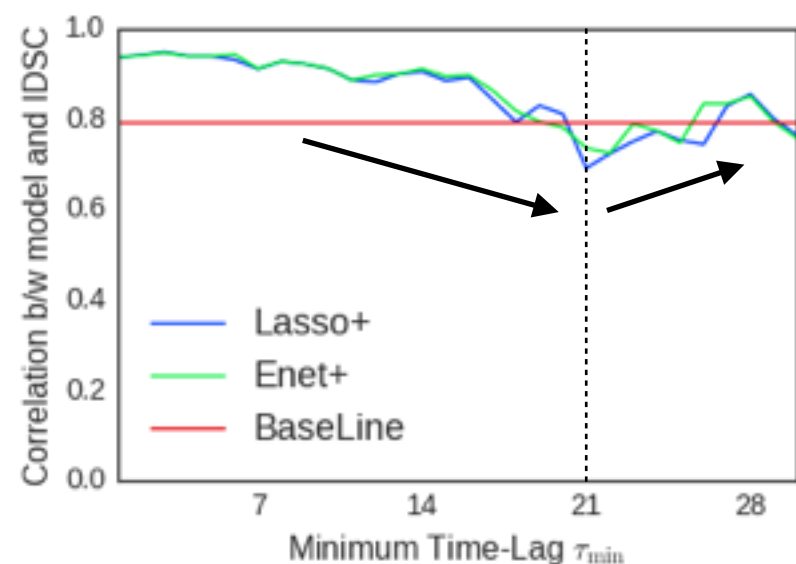
予測精度

考察

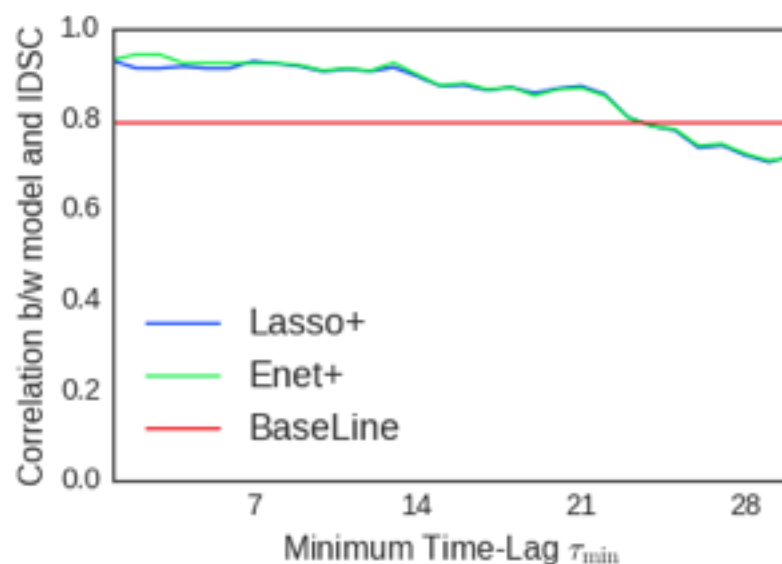
問題点とその改善法

予測精度が非定常な振る舞い

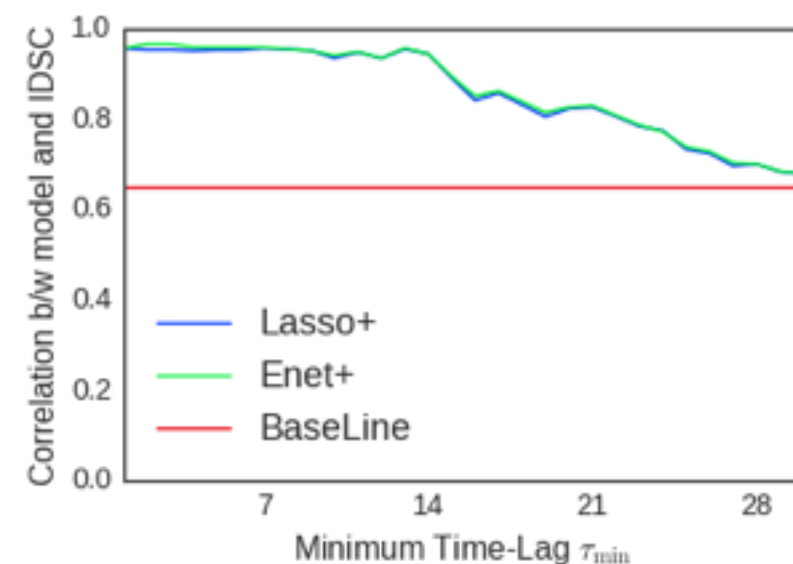
Train: Season 2/ Test: Season 1



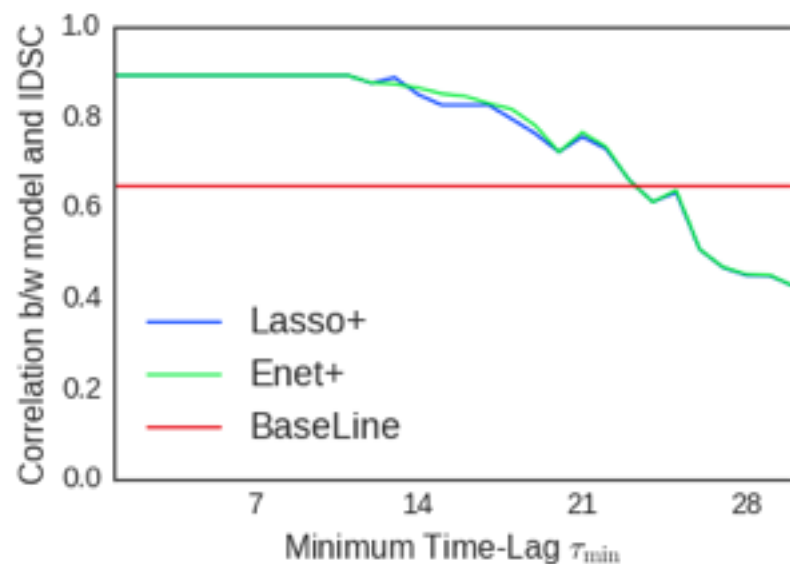
Train: Season 1/ Test: Season 2



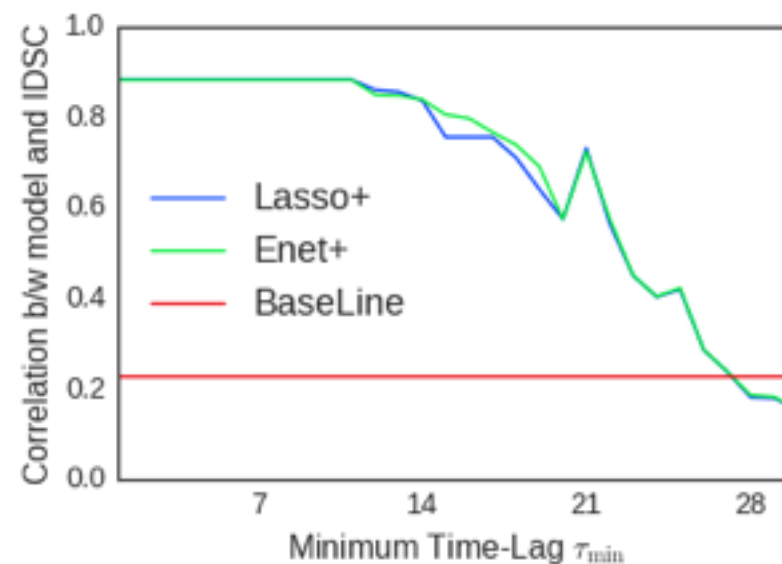
Train: Season 1/ Test: Season 3



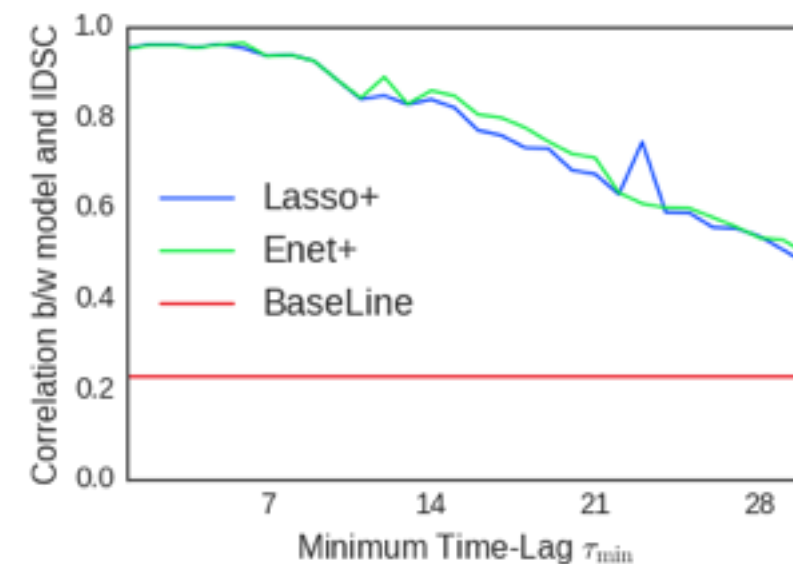
Train: Season 3/ Test: Season 1



Train: Season 3/ Test: Season 2

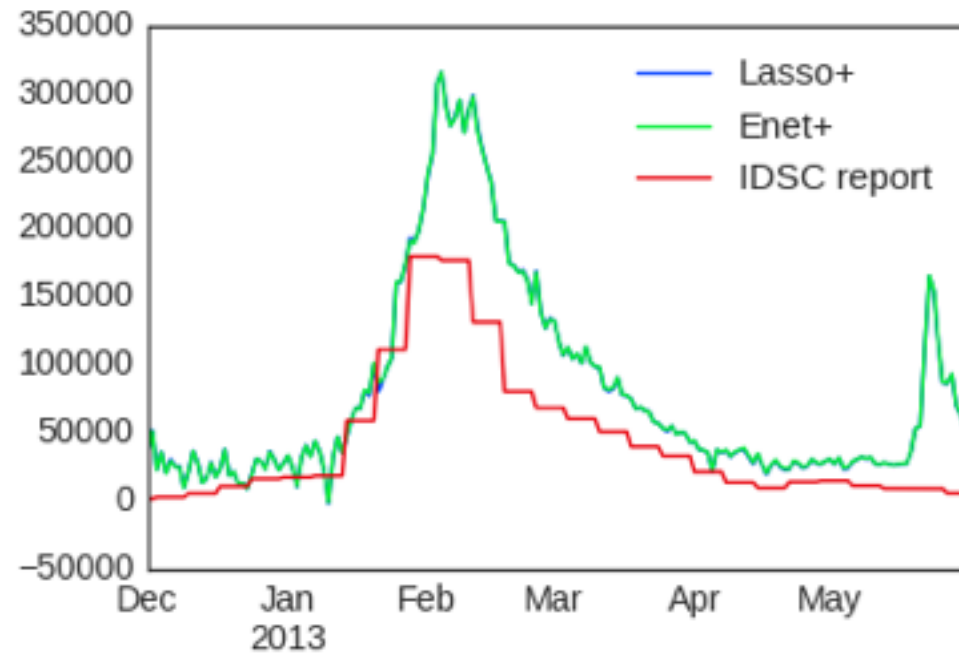


Train: Season 2/ Test: Season 3

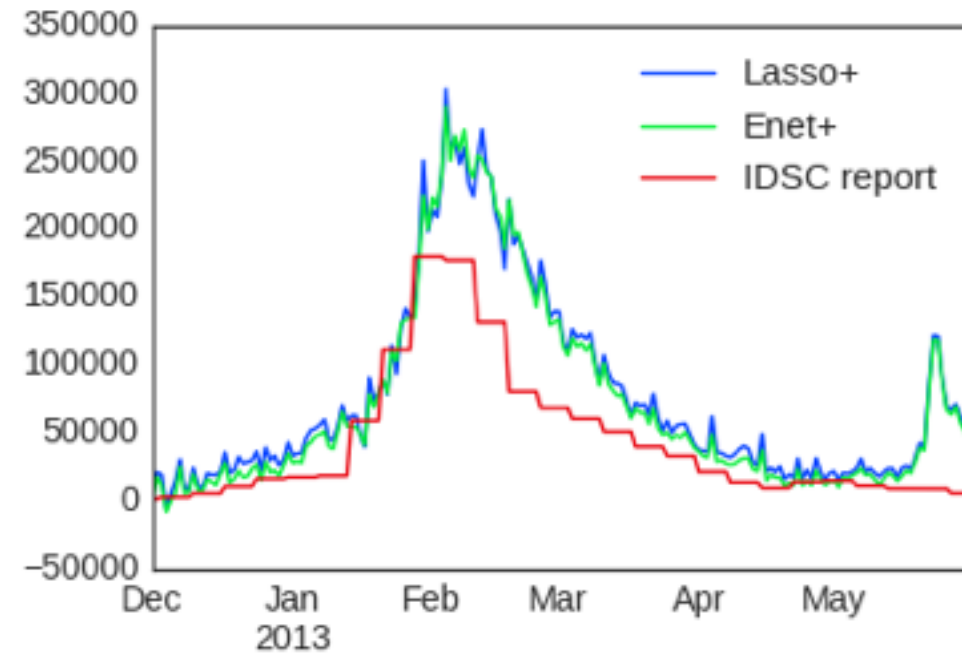


終盤に患者数を過剰推定

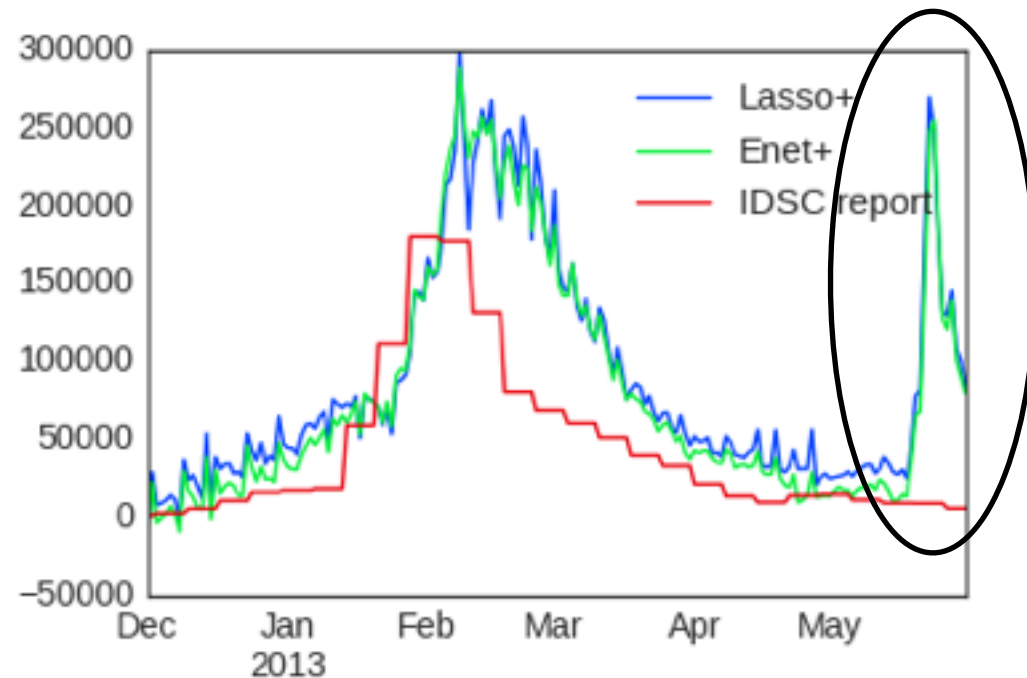
$\tau_{\min}$: 1 week



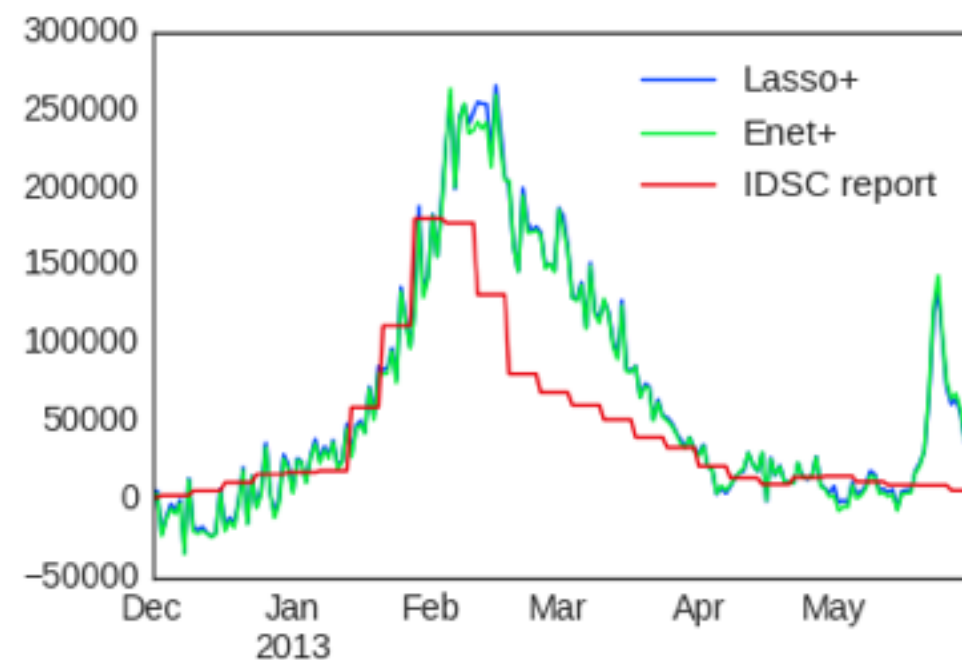
$\tau_{\min}$: 2 week



$\tau_{\min}$: 3 week

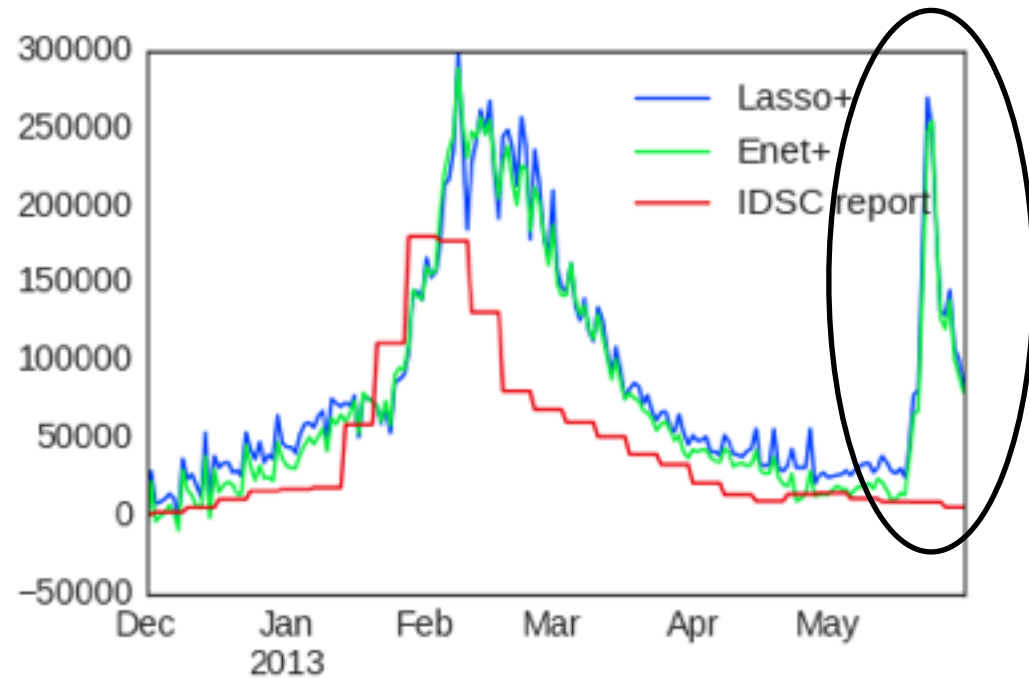


$\tau_{\min}$: 4 week

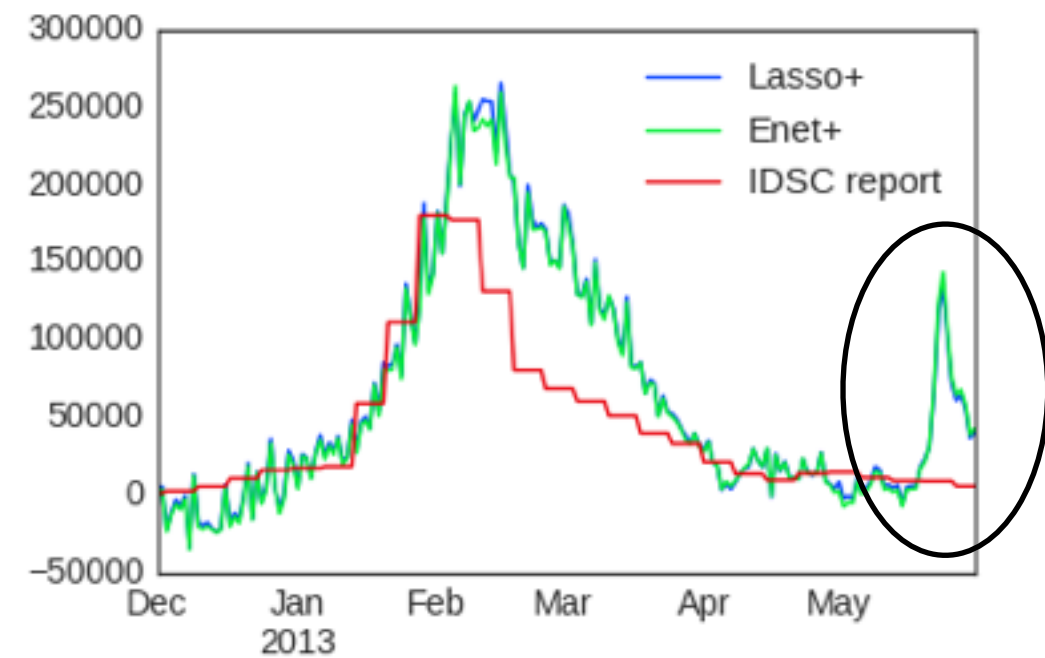


4 week の時には重み 減 -> 予測精度が改善

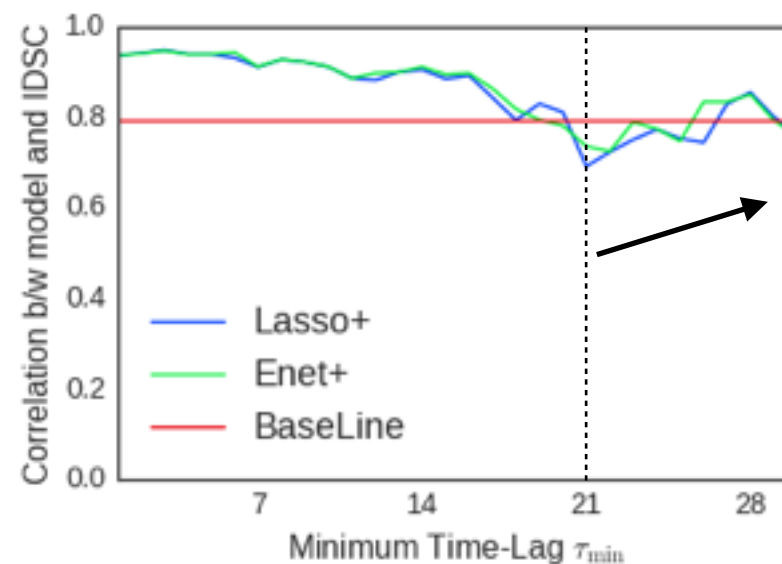
$\tau_{\min}$: 3 week



$\tau_{\min}$: 4 week

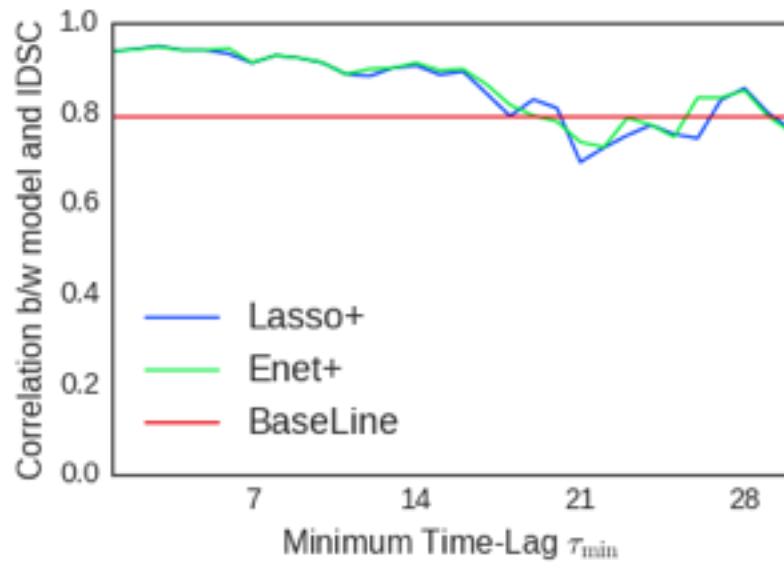


Train: Season 2/ Test: Season 1

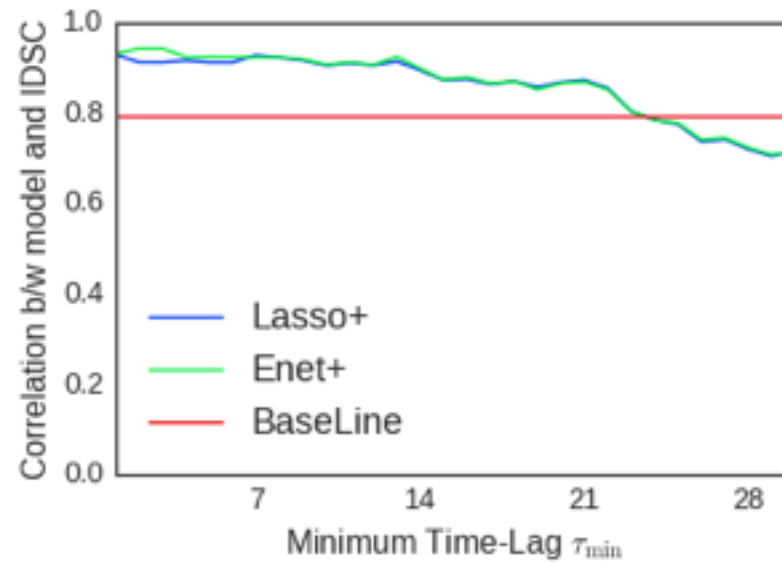


推定精度が最も悪いケース

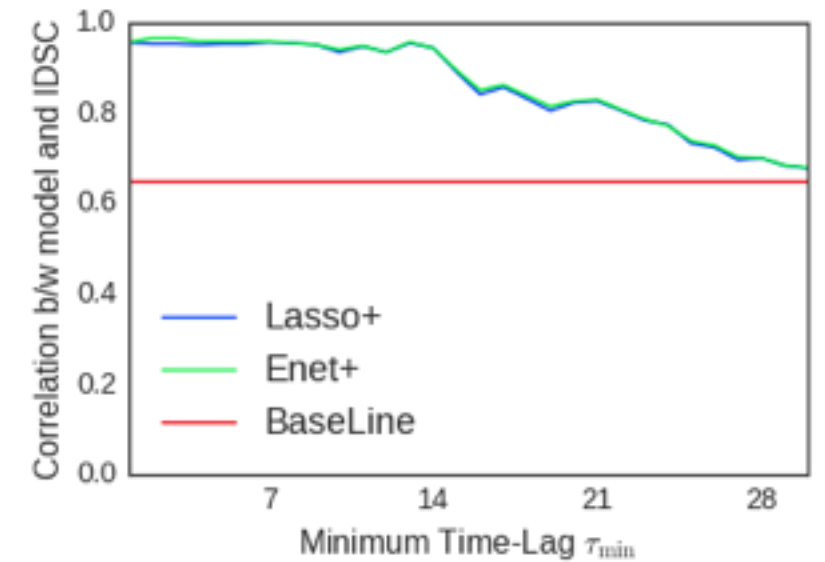
Train: Season 2/ Test: Season 1



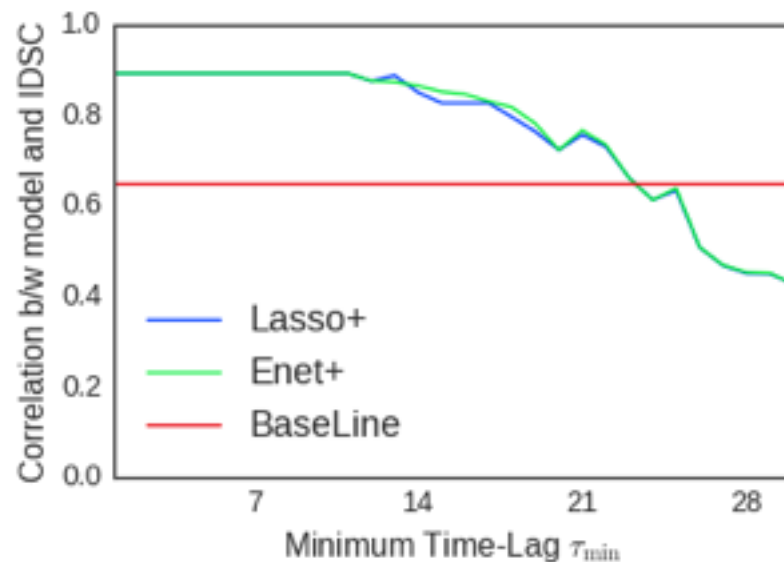
Train: Season 1/ Test: Season 2



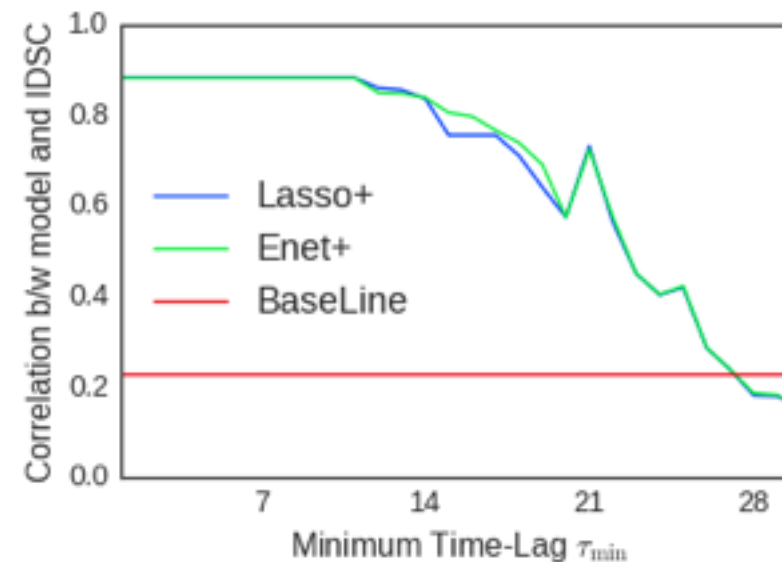
Train: Season 1/ Test: Season 3



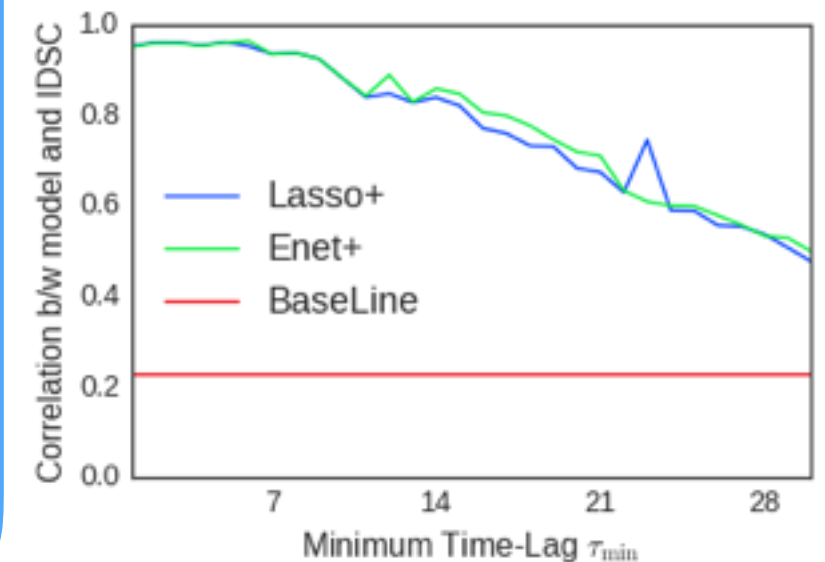
Train: Season 3/ Test: Season 1



Train: Season 3/ Test: Season 2

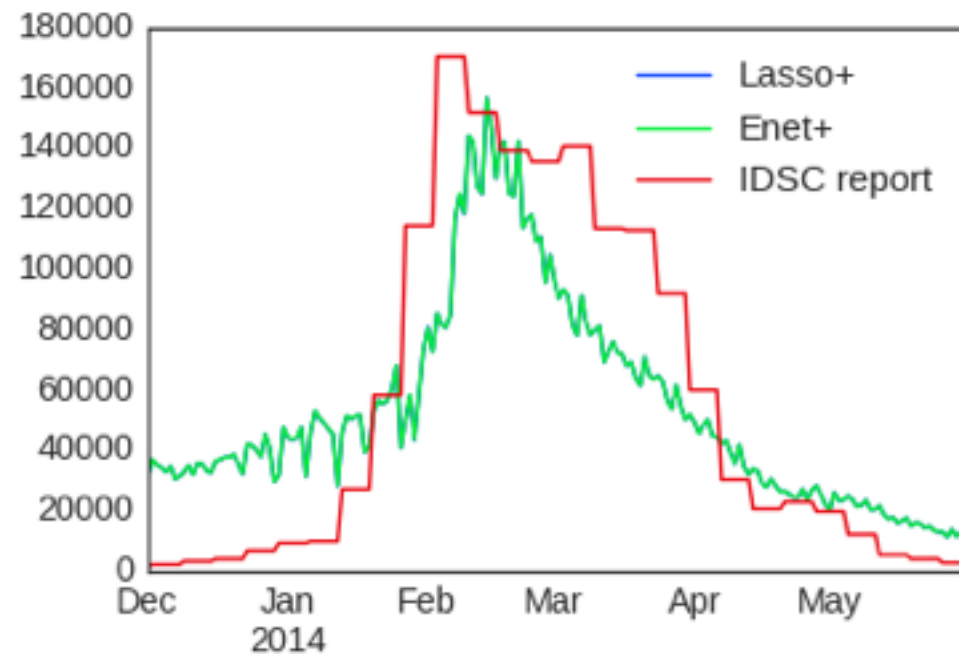


Train: Season 2/ Test: Season 3

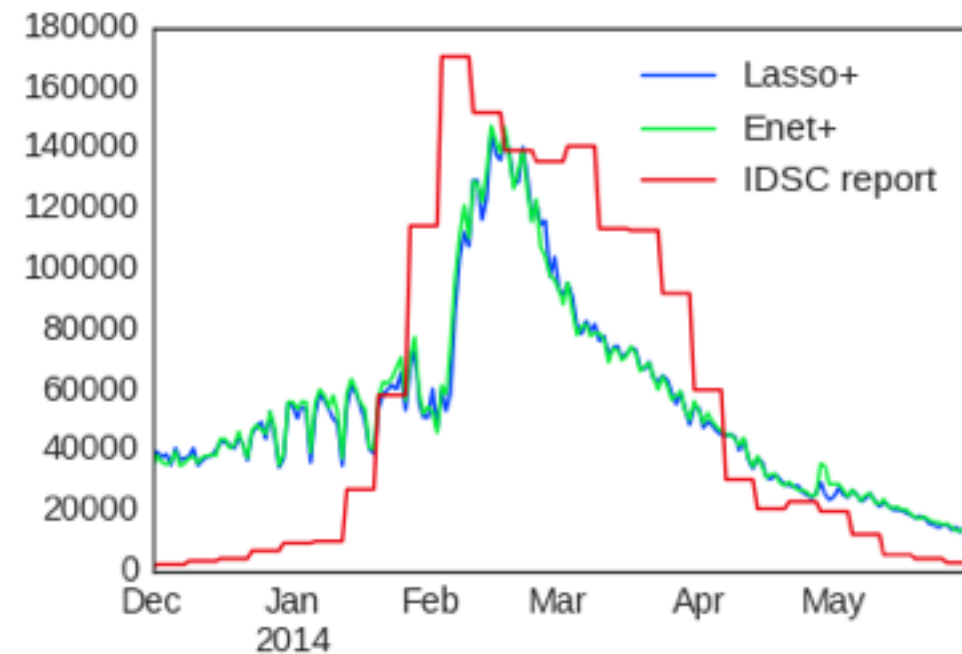


実際の流行前に患者数を過剰推定

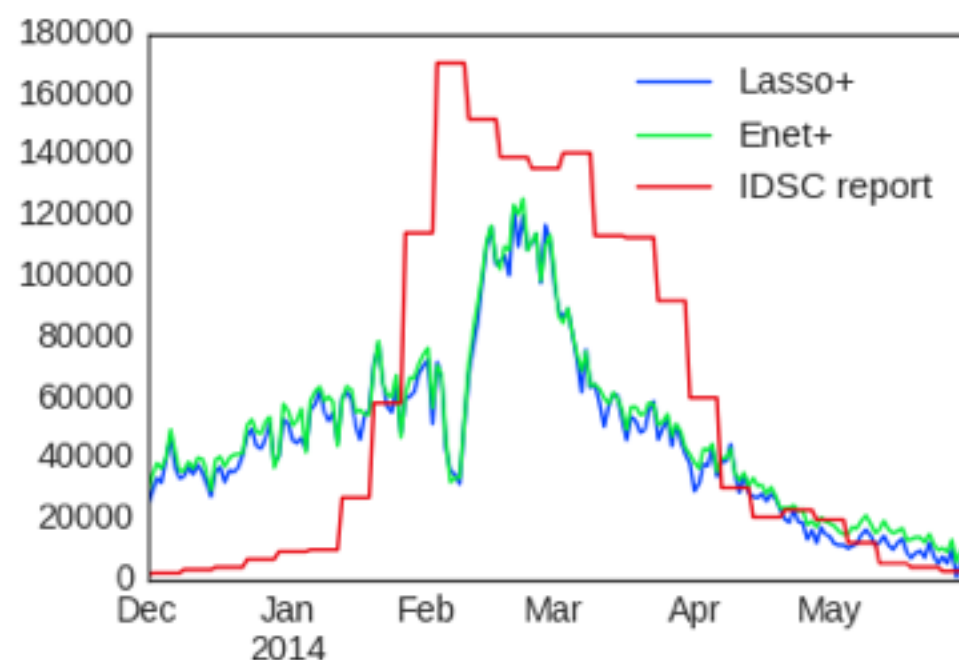
$\tau_{\min} : 1 \text{ week}$



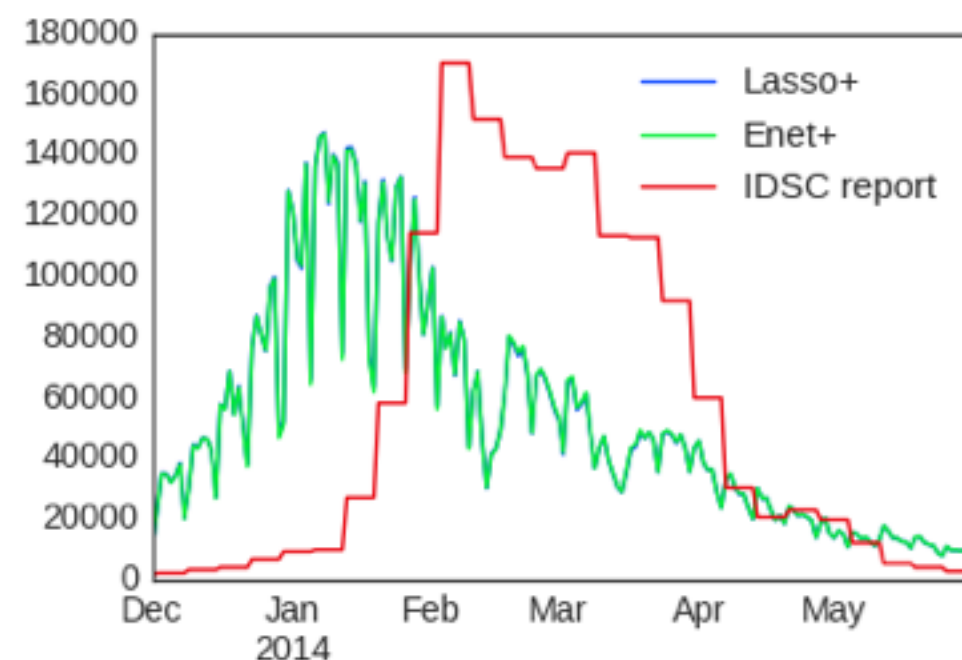
$\tau_{\min} : 2 \text{ week}$



$\tau_{\min} : 3 \text{ week}$

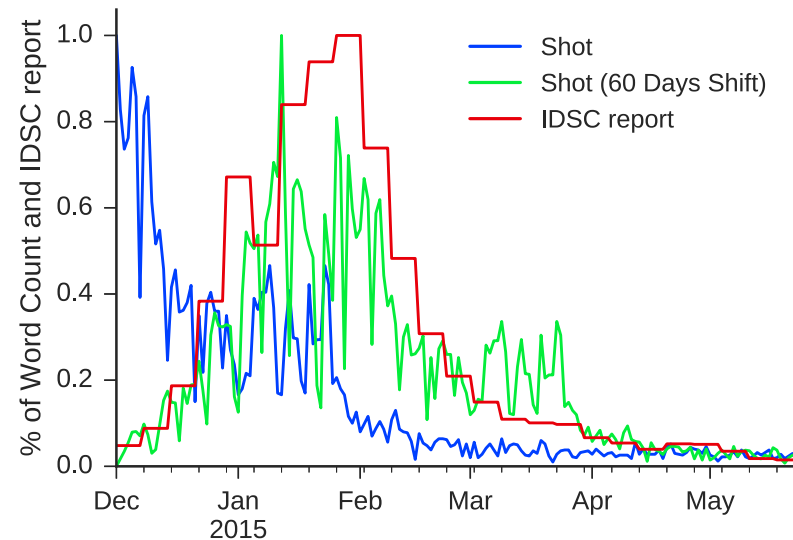


$\tau_{\min} : 4 \text{ week}$

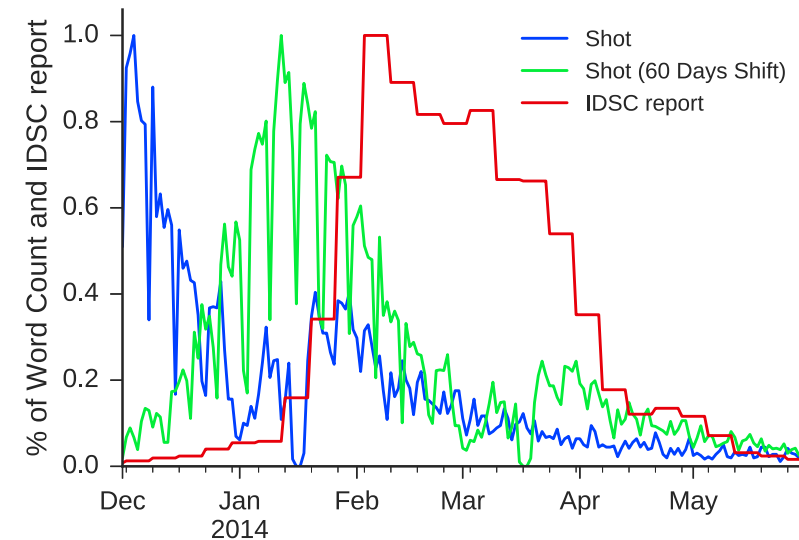


原因: 周期性のある語”打つ”を選択

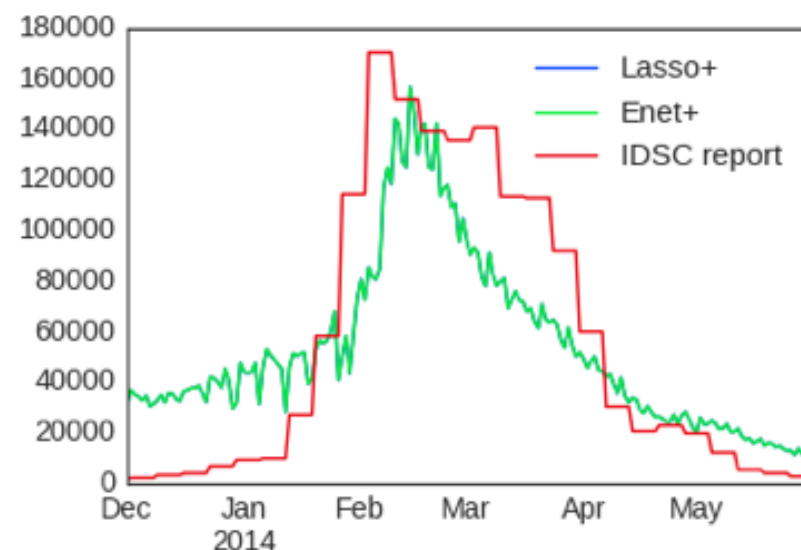
Train



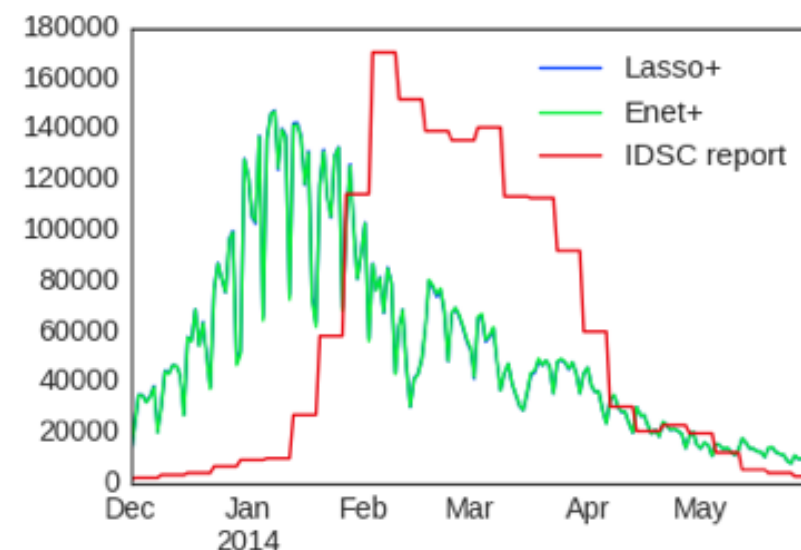
Test



$\tau_{\min}$: 1 week



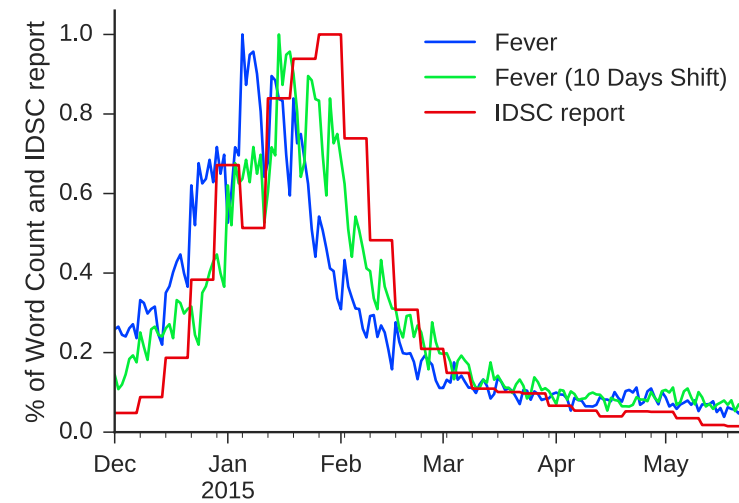
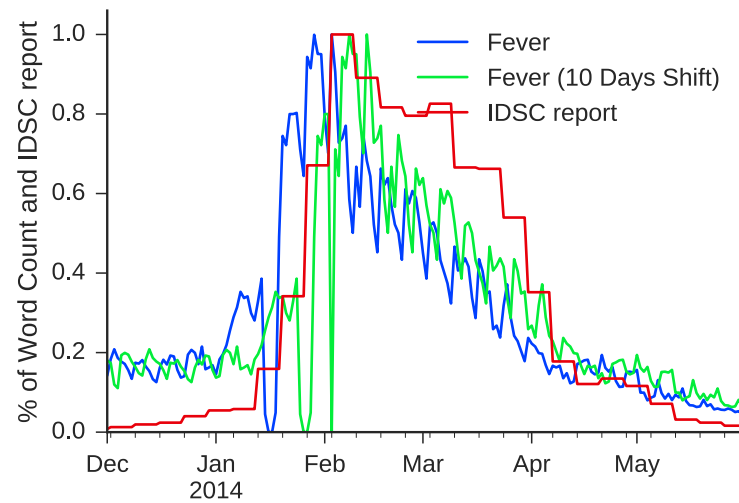
$\tau_{\min}$: 4 week



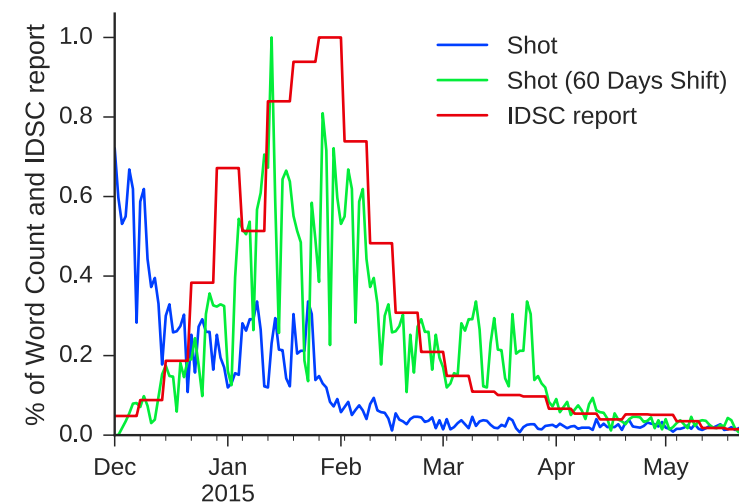
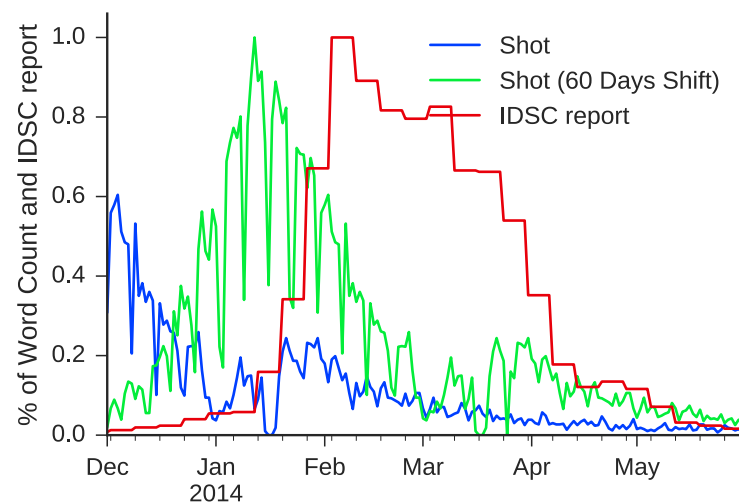
$\tau_{\min}$ 大 $\longrightarrow$ β_{Shot} 大

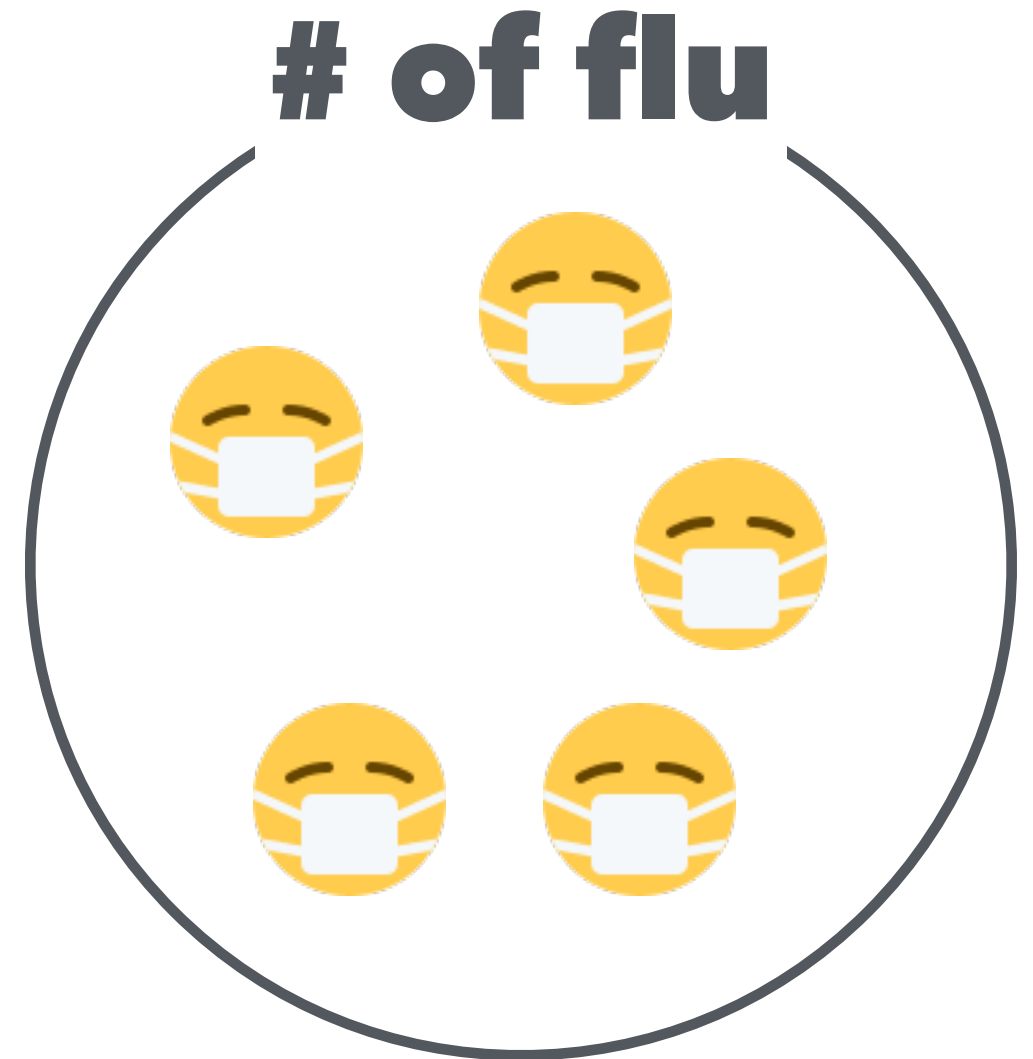
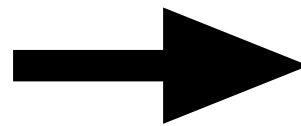
解決案: 2年以上のデータから学習

一貫した傾向: 相互相関 大



異なる傾向: 相互相関 小





Summary

感染症サーベイランスの重要性と本研究のまとめ

Summary

* インフルエンザにより世界で毎年**25万 ~ 50万人が死亡***

- ・ 感染の広がりが予測可能となり早期対策を打てる。

(今回の成果を共同研究先の国立感染症研究所に提供予定)

* 単純な手法で高い精度の実現

- ・ 本邦**最高精度**かつ精度向上の余地が多く残されている

- ・ **コーパス(tweetID)公開**予定

- ・ NTCIR13-MedWeb (2017)

* **広く適用可能**な手法 (タイムラグのある関係を取り込める)

Thank you for listening 🤗

背景

インフルエンザサーベイランスの重要性

手法

タイムラグを扱う

予測への拡張

実験

推定精度

予測精度

考察

問題点とその改善法