リーディング DAT LB1　レポート

2018/02/27　鈴木　毅洋

※本レポート作成にあたり使用したコードは下記にアップした。
https://github.com/statefb/KFAS_practice/blob/master/practice/report.ipynb

データ概要

2003 年 1 月から 2016 年 12 月までの訪日外客数(以降，来日数)を示す。

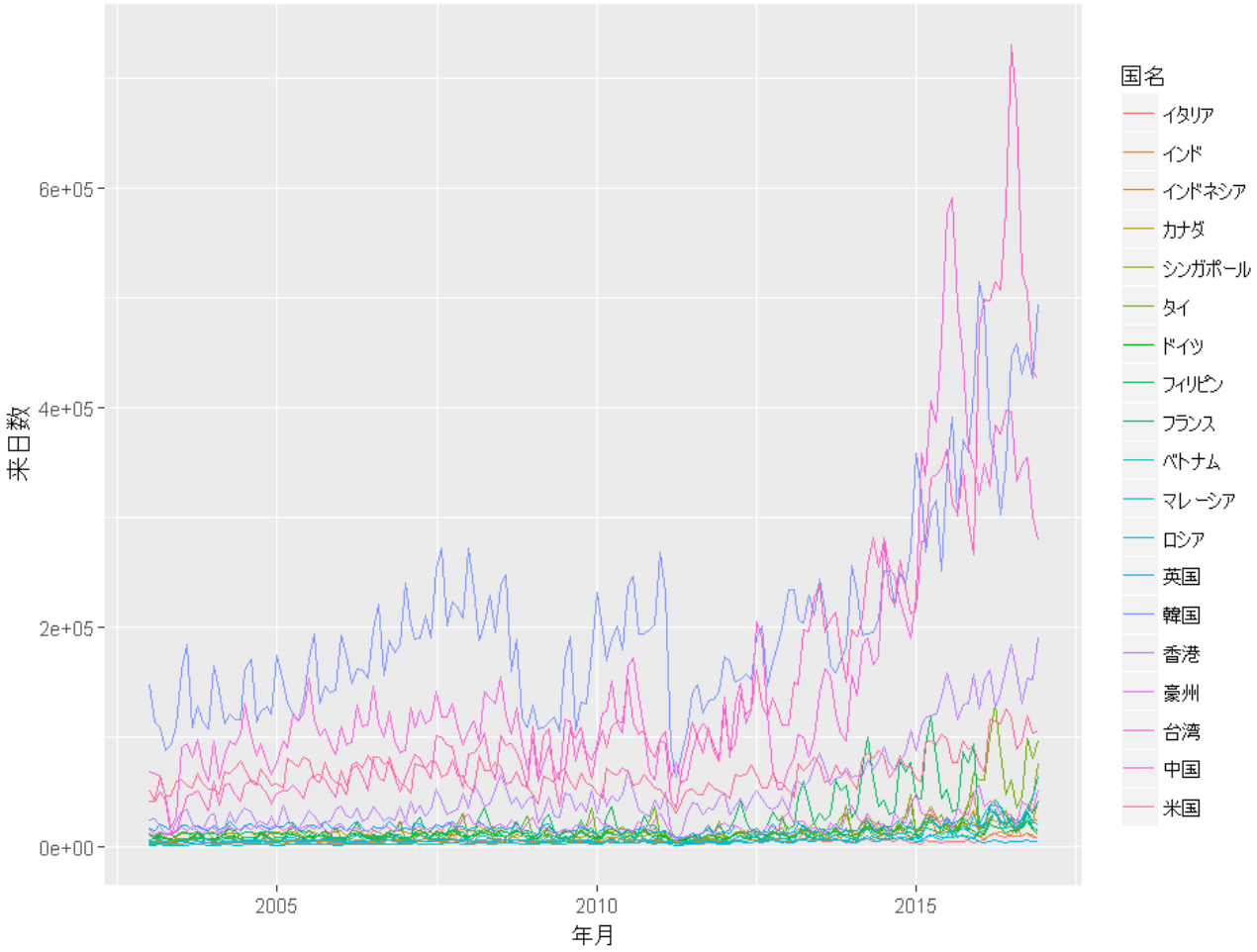

図 1　来日数

図 1 より下記のことがわかる。

- 以下のイベントに対して，トレンドの大きな落ち込みが見られる。
  - 2008 年：リーマンショック
  - 2011 年：東日本大震災
- 12 ヶ月を 1 周期とする周期性が見られる。変動幅は来日数の規模に応じて変化している。

モデル作成

本レポートでは，予測対象国は米国とする（選定理由は特にない）。
まずは幾つかのモデルを作成し，その中で AIC 最小モデルの結果をプロットしてみる。

モデルの作成方針を下記に示す。

- 周期性が見られるため、季節変動成分をモデルに加える。
- 2012 年以降、トレンドは上昇傾向にあるため、ローカルレベル＋平滑化トレンドモデルを採用する。
- ここではまず，説明変数はカレンダーのみ考慮する。
- 観測値はカウント値で上限がないため、観測値がポアソン分布から生成されるモデルが考えられる。
  しかしトレンドグラフより平均≠分散であることが伺えるため、観測値の対数が正規分布に従うモデルを採用する。

図 1 の対数をとったトレンドを図 2 に，米国のみの対数トレンドを図 3 にそれぞれ示す。

図 2 来日数(対数)

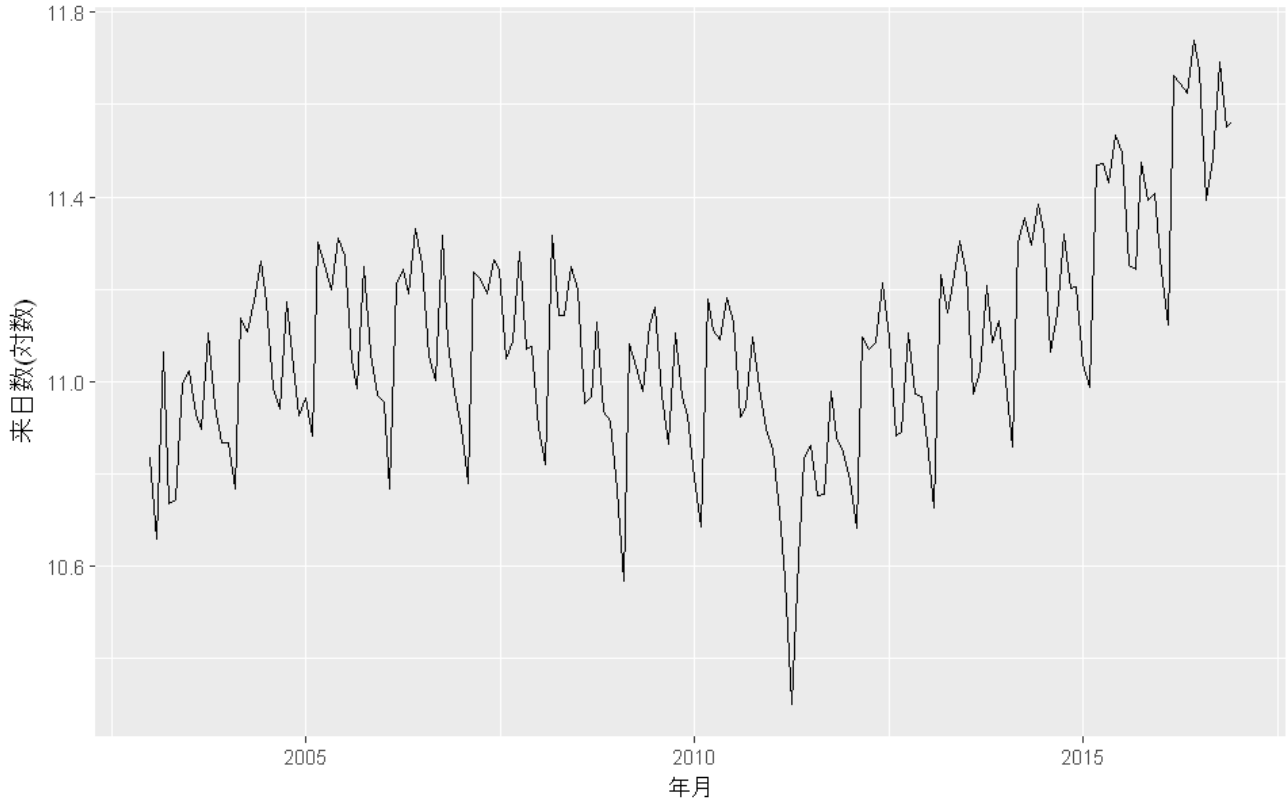

図 3 米国人の来日数(対数)

米国のトレンドを見ると，震災の影響こそ他国と同様にみられるものの，リーマンショックの影響は比較的小さいように見える。

上述した方針に基づき，下記 6 つのモデルを比較する。

- モデル 1：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動なし)
- モデル 2：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)
- モデル 3：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動あり)
- モデル 4：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動あり)
- モデル 5：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動なし)＋カレンダー成分
- モデル 6：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)＋カレンダー成分

算出された AIC 及び対数尤度を下記図 4 に示す。

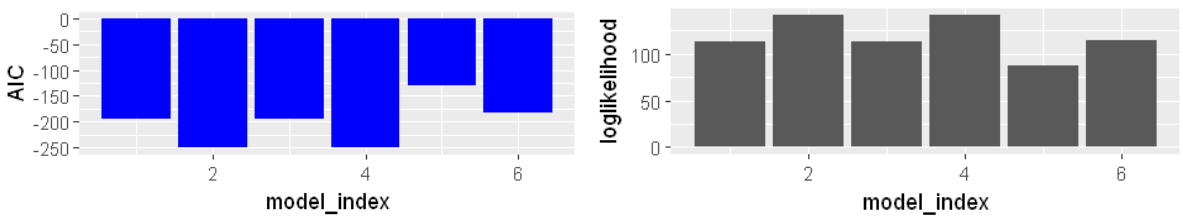

図 4 AIC(左)，対数尤度(右)

図 4 より，モデル 2，4 が良いことがわかる。カレンダー効果を考慮したモデル 5，6 は対数尤度，AIC 共に悪化しており，カレンダーの影響は考慮しない方が良いことが分かる。

最も AIC の小さなモデル 2 の結果を以下に示す。

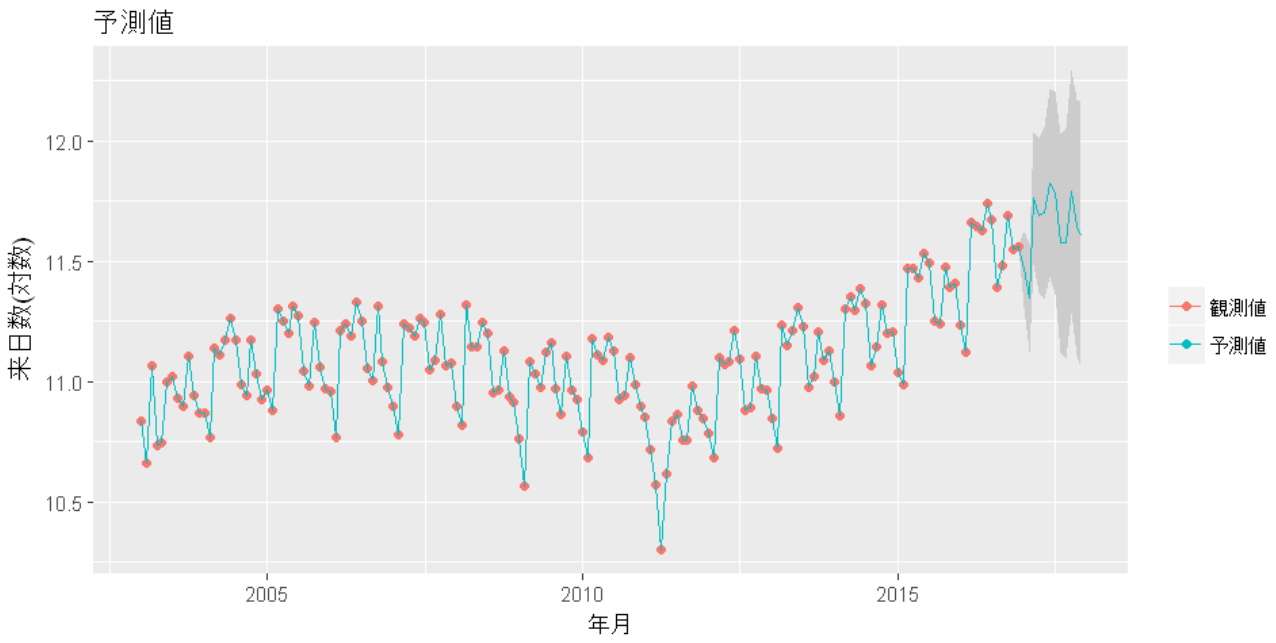

図 5 予測結果(網掛けは予測値 95%信頼区間)

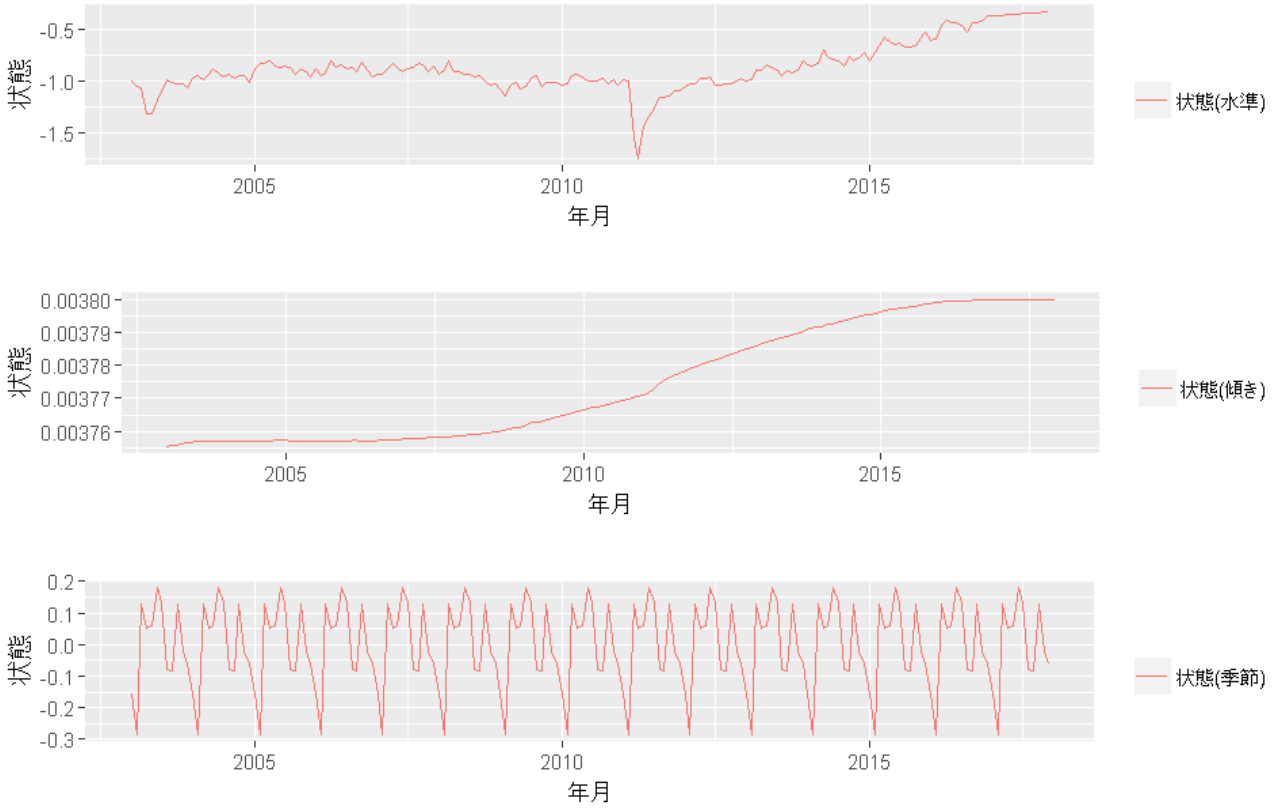

図 6 状態の推移

図 7 残差

図 6 の水準の推移を見ると，震災の付近で 0.7 以上の大きな落ち込みが確認できる。推定された観測誤差分散と水準誤差分散はそれぞれ 5.7e-7，6.2e-3 であり，この落ち込みは特に当てはまりが悪いサンプルであると言える。

図 7 の残差を見ると，2011 年 2 月～同年 4 月において大きな外れ値が確認でき，残差解析からも震災の影響を大きく受けていることが分かる。一方でリーマンショックの影響は小さく，米国に関しては考慮する必要はないと言える。

外れ値除去

この結果を踏まえ，震災付近のデータ（2011 年 2 月～同年 4 月）を外れ値とみなして除外し，説明変数に震災前か後かを表す変数（以降，地震変数）を新たに加えて再度解析を行う。地震変数は 2011 年 2 月以前を 0，3 月以降を 1 とするカテゴリカル変数である。

本レポートでは以下 12 個のモデルの比較を行った。

- モデル 1：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動なし)
- モデル 2：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)
- モデル 3：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動あり)
- モデル 4：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動あり)
- モデル 5：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動なし)＋カレンダー成分
- モデル 6：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)＋カレンダー成分
- モデル 7：滑化トレンド(水準変動なし)＋周期性(季節変動なし)＋地震
- モデル 8：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)＋地震
- モデル 9：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動あり)＋地震
- モデル 10：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動あり)＋地震
- モデル 11：ローカルレベル＋平滑化トレンド(水準変動なし)＋周期性(季節変動なし)＋地震＋カレンダー成分
- モデル 12：ローカルレベル＋平滑化トレンド(水準変動あり)＋周期性(季節変動なし)＋地震＋カレンダー成分

算出された AIC 及び対数尤度を下記図 8 に示す。

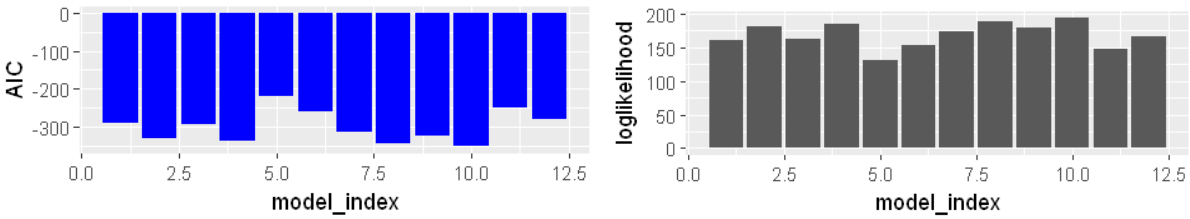

図 8 AIC(左)，対数尤度(右) ※外れ値除去後

図 8 を見ると，モデル 10・モデル 7（地震変数あり）が特に良い結果となっている。これらのモデルは，外れ値除外前で選択されたモデルに地震の影響を考慮したものである。またカレンダー効果は対数尤度，AIC を悪化させているが，これは外れ値除去前と同様である。

ここでは AIC が最も小さなモデル 10 の結果をプロットする。

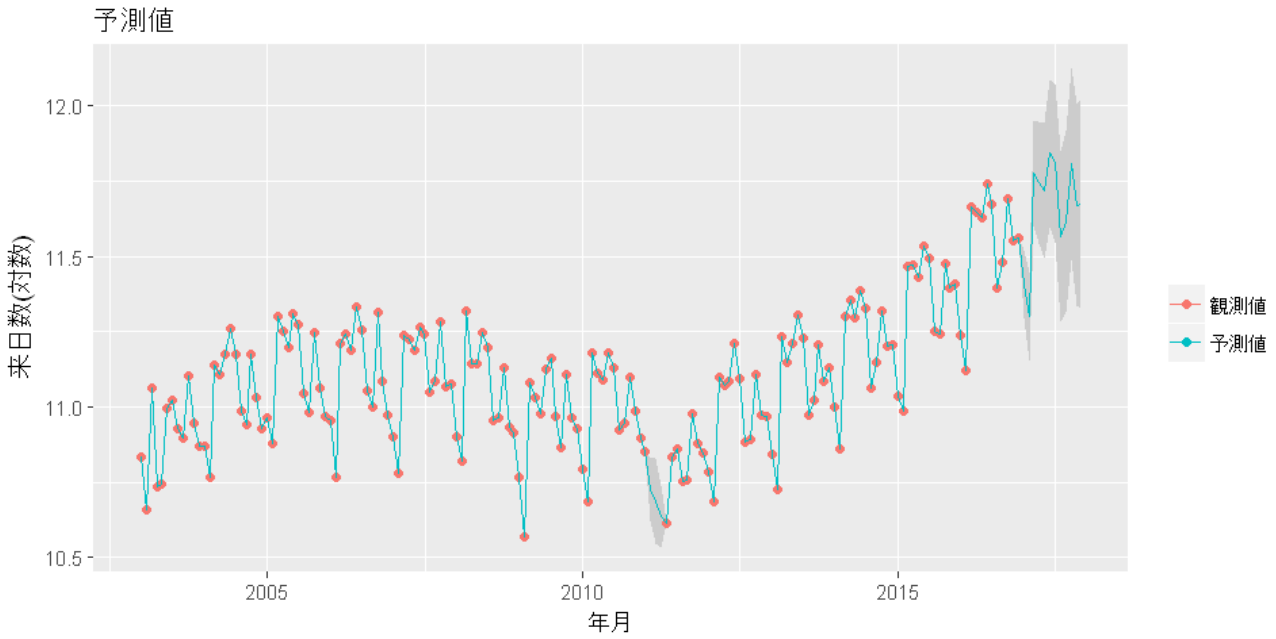

図 9 予測結果(網掛けは予測値 95%信頼区間) ※外れ値除去後

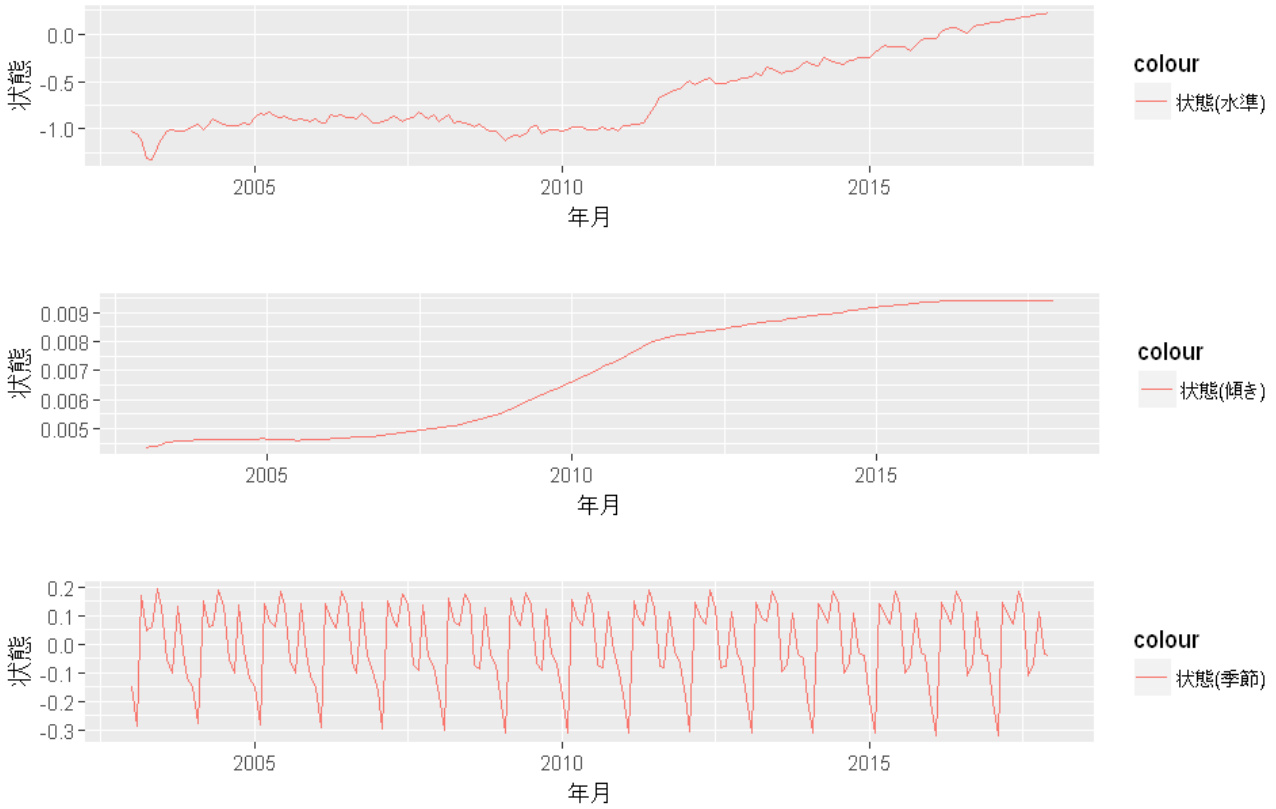

図 10 状態の推移 ※外れ値除去後

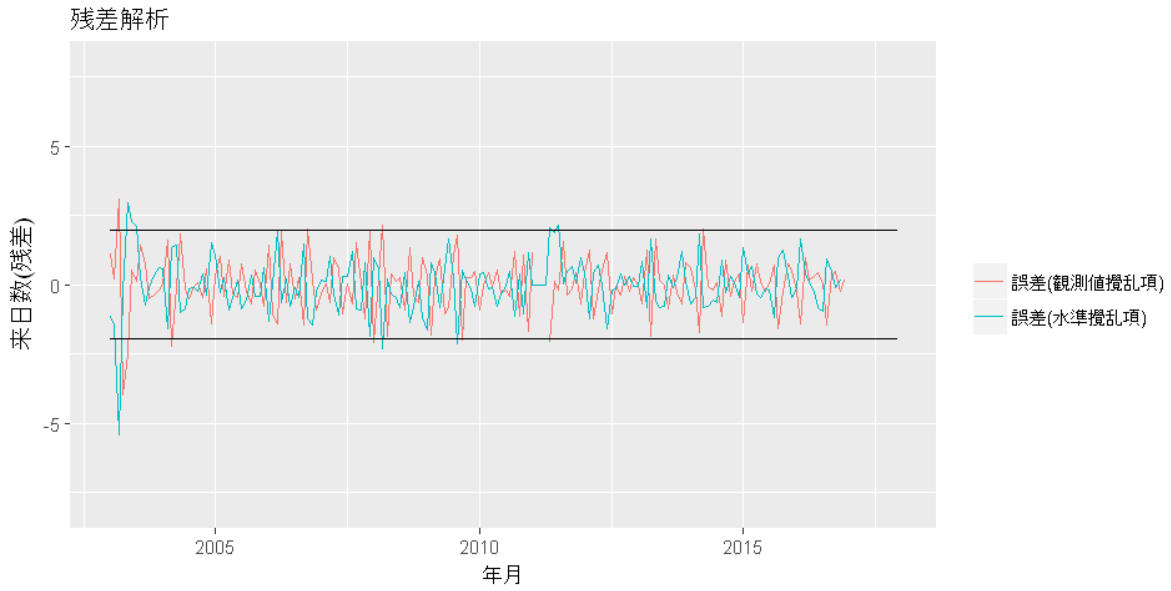

図 11 残差 ※外れ値除去後

図 10 の水準の遷移を見ると，外れ値除去前と比較し滑らかに推移している。また図 11 の残差解析においても，震災付近前後で 95%区間を大きく逸脱しているサンプルは見られなくなった。以上のことから，外れ値を除去したことで，状態遷移および観測誤差にガウス分布を仮定する本モデルにおいて，より妥当な結果が得られたと言える。

予測結果

以上の解析から，2017 年来日数の予測に，外れ値除去後のモデル 10 を採用する。これまで対数スケールで分析を実施してきたため，元のスケールに戻した結果を図 12 に示す。また予測値の期待値を表 1 に示す。

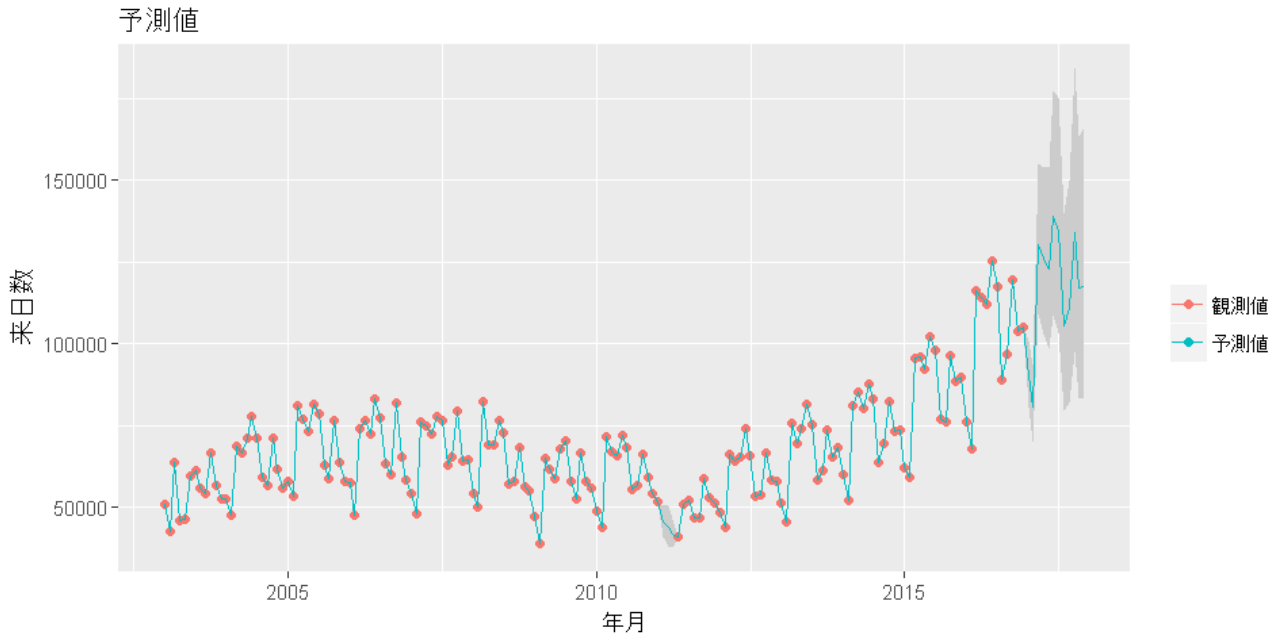

図 12 予測結果(元のスケール)

表 1 2017 年　米国人の予想来日数（単位：千人）

1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
90.0	80.8	130	126	123	139	135	105	111	134	117	118

番外編：bsts パッケージによる分析

上述までの結果は KFAS パッケージによる分析結果である。本レポートではさらに，ベイジアン構造時系列 (Bayesian structual time series) を扱うことのできる，bsts パッケージを使った分析を実施したので掲載する。本分析においては下記 URL 記事を参考にした。

[R] bsts パッケージの使い方：http://ill-identified.hatenablog.com/entry/2017/09/08/001002

KFAS パッケージではパラメータの推定を最尤推定によって実施するが, bsts パッケージは事前分布を設定し，サンプリングによりベイズ推定する点が異なる。また非ガウス分布のモデルもある程度扱うことが可能である。

bsts パッケージの特筆すべき事として, lasso のように, 説明変数選択の機能が内臓されている点が挙げられる。各説明変数の係数に spike-and-slab 分布を仮定することで，スパースな結果を得ることができる。本レポートの題材は説明変数が少ないため効果は薄いが，大量にあるケースでは有用であると考えられる。

KFAS の分析時と同様に，まずは外れ値を除去しないケースで分析を行う。
ここでは下記 2 つのモデルを比較する。なお，状態遷移誤差および観測誤差はガウス分布に従うものとし，データは対数変換を施し標準化を行ったものを対象とした。

- モデル 1：ローカルレベル＋平滑化トレンド＋季節変動
- モデル 2：ローカルレベル＋平滑化トレンド＋季節変動＋説明変数（カレンダー効果のみ）

推定されたモデルそれぞれにおいて，1 期先予測値の累積誤差を以下図 13 に示す。

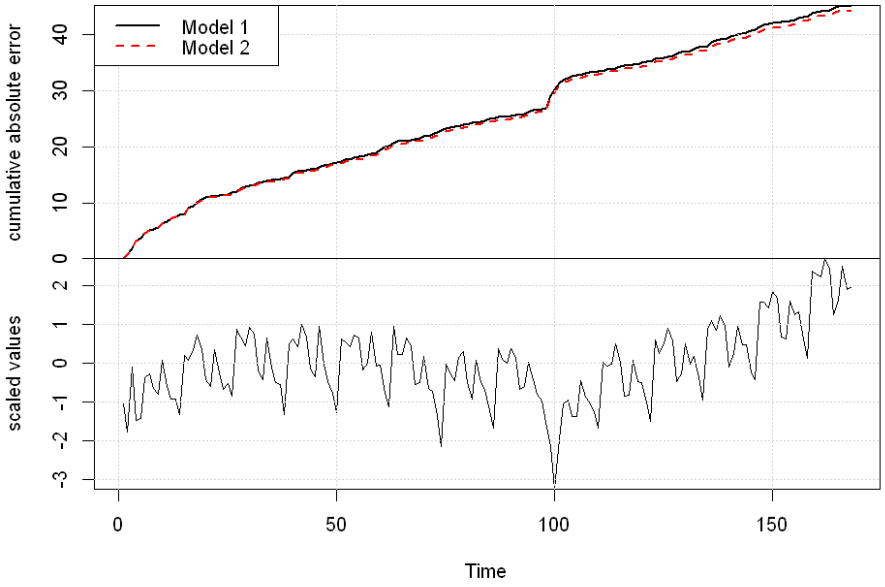

図 13 累積誤差の比較 (横軸：データのインデックス)

震災後に誤差が大きく上昇し，大震災の影響が無視できないことがわかる。
次に外れ値を除去し，地震変数を加えたモデルで分析を行う。比較したモデルは下記のとおりである。

- モデル 1：ローカルレベル＋平滑化トレンド＋季節変動
- モデル 2：ローカルレベル＋平滑化トレンド＋季節変動＋説明変数（カレンダー効果＋地震）

推定されたモデルそれぞれにおいて，1 期先予測値の累積誤差を以下図 14 に示す。

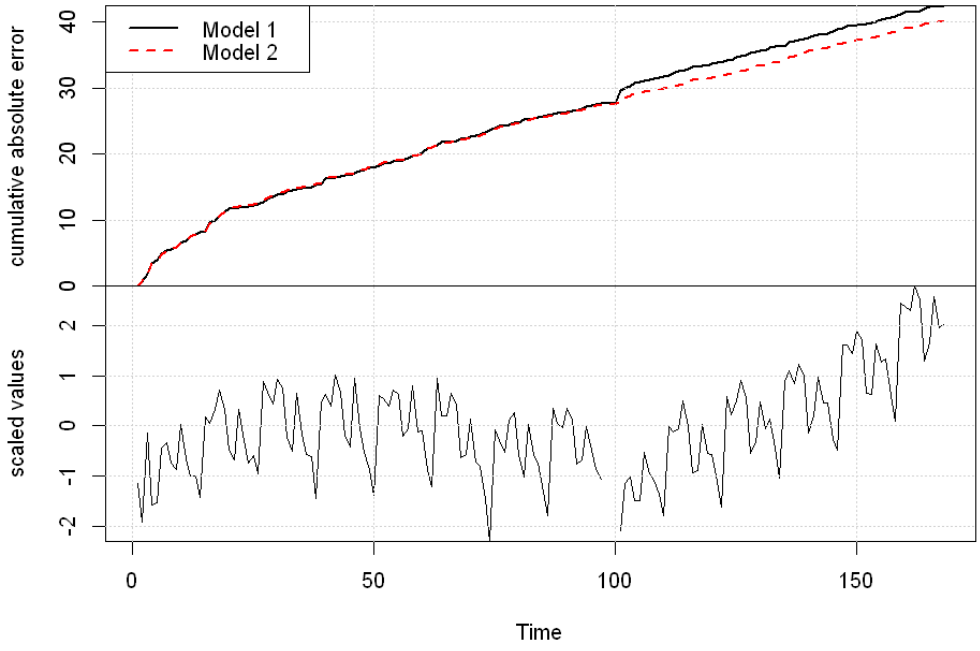

図 14 累積誤差の比較 (横軸：データのインデックス)　※外れ値除去後

外れ値の除去により、累積誤差が改善されたことがわかる。また震災前では、２つのモデルに大きな差は見られない一方で，震災後の累積誤差を見ると，モデル 1 は震災直後にステップ状に累積誤差が増加している。これは震災直後に大きな傾向変化があったことを表している。

モデル 2 の各説明変数がモデルに含まれる確率を以下図 15 に示す。

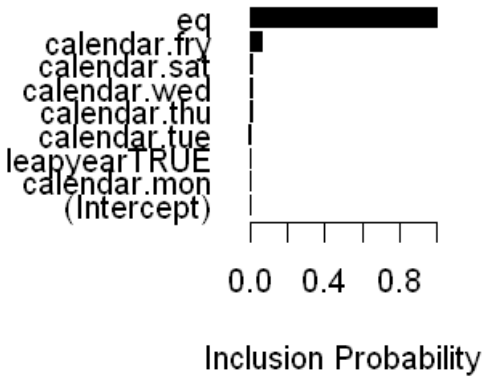

図 15 モデル 2 の説明変数含有確率 (eq：地震，calendar.＊：曜日効果，leapyear：うるう年効果)

図 15 より，地震の変数はほぼ 10 割選択されている一方で，カレンダー効果はほとんどないことがわかる。これは KFAS の分析結果と同等の結果である。