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。

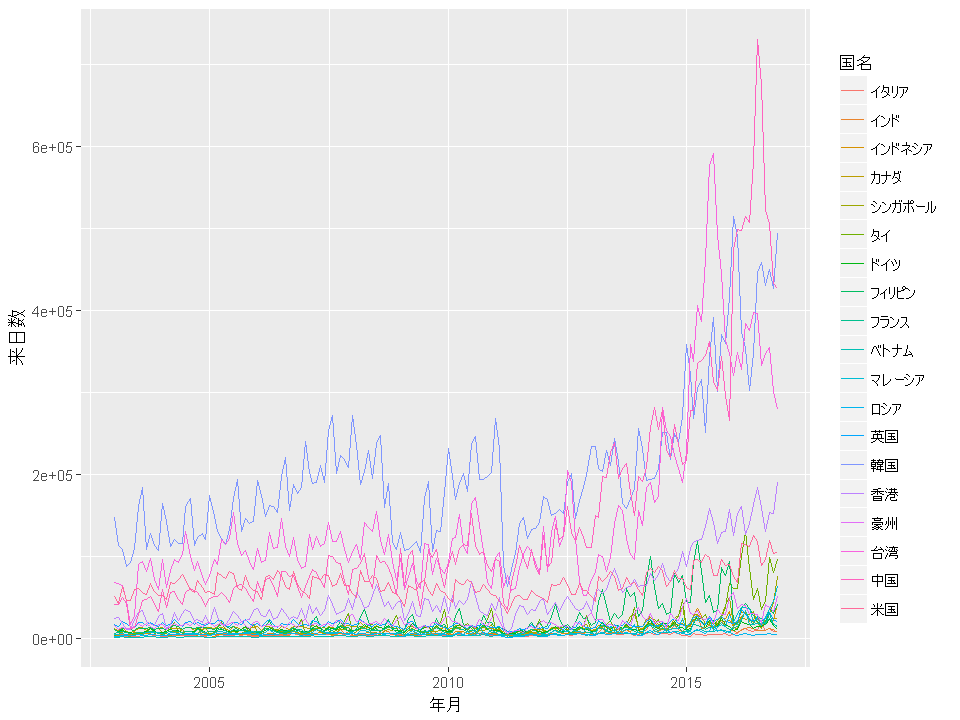


図1 来日数

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。

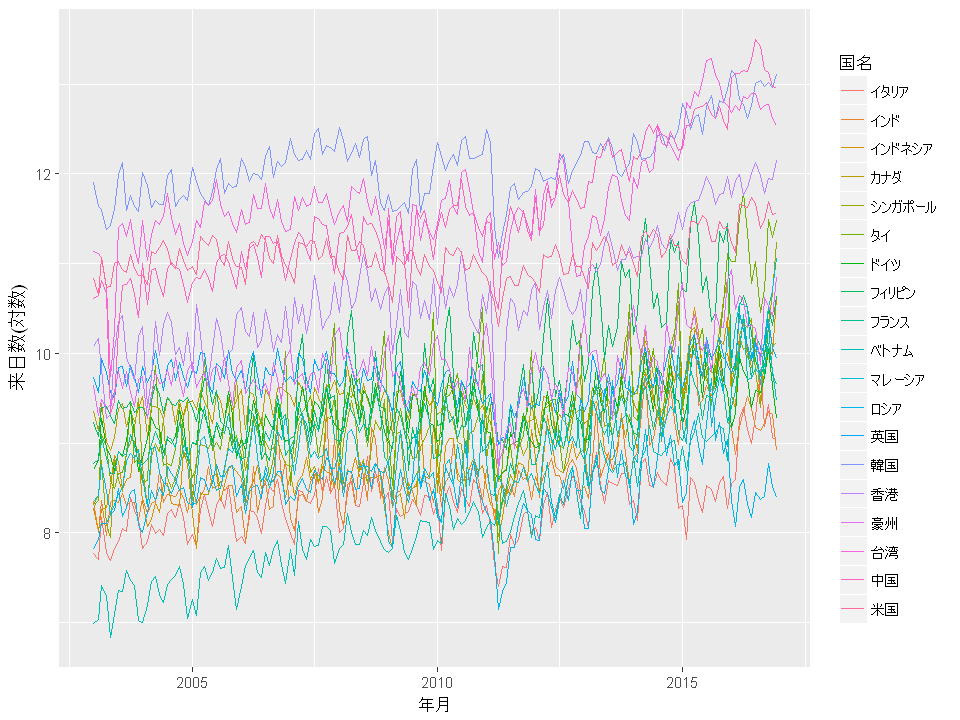


図2 来日数(対数)

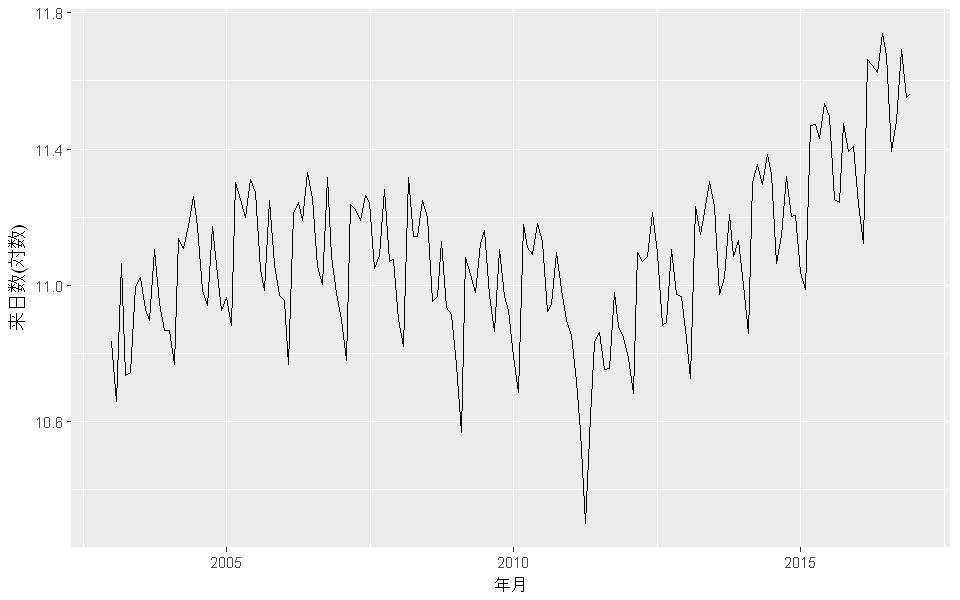


図3 米国人の来日数(対数)

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。

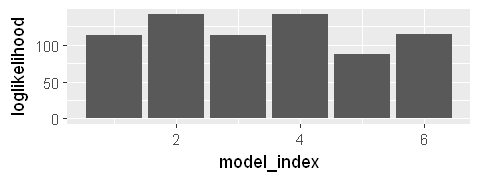
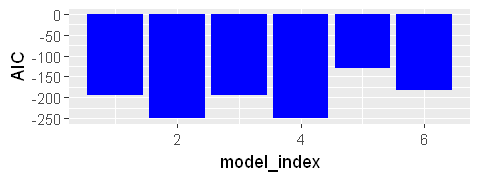


図4 AIC(左)，対数尤度(右)

　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AICが悪化しており，カレンダーの影響は薄いと考えられる。

最もAICの小さなモデル2の結果を以下に示す。

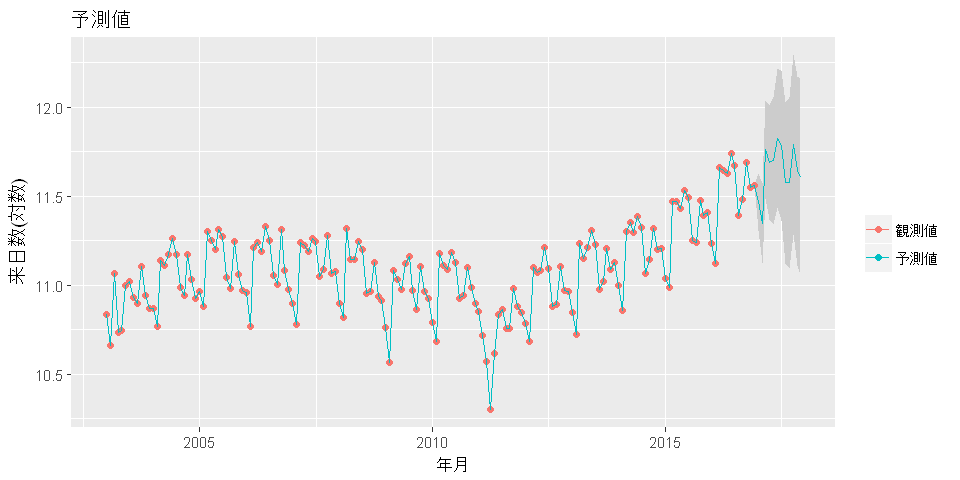
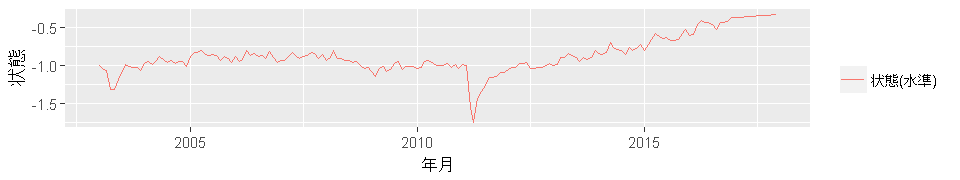
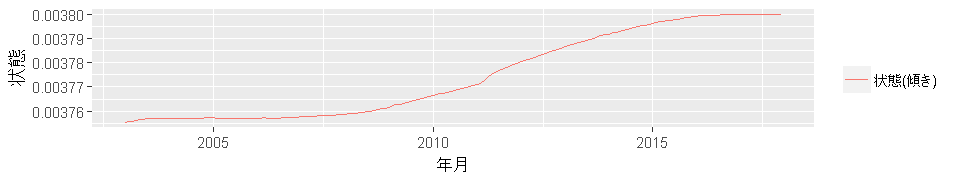


図5 予測結果(網掛けは予測値95％信頼区間)





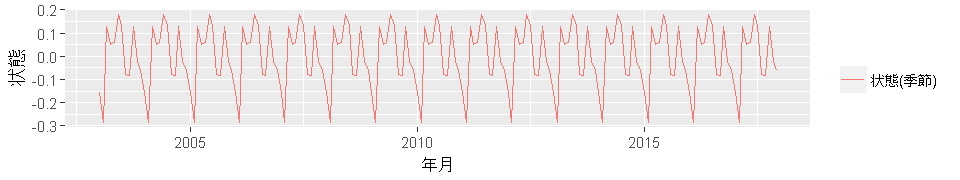


図6状態の推移

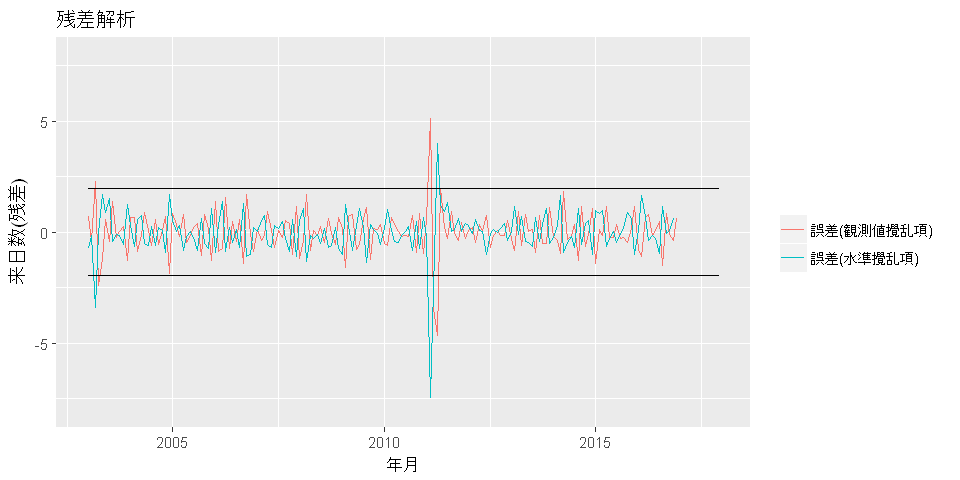


図7 残差

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。

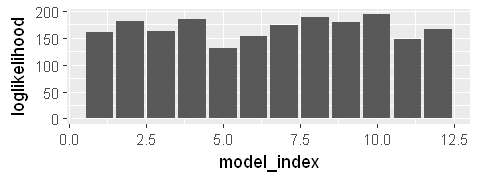
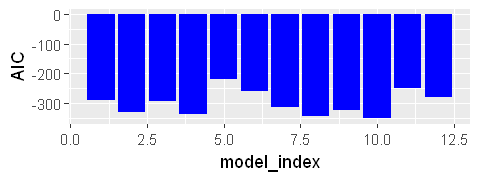


図8 AIC(左)，対数尤度(右) ※外れ値除去後

　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。

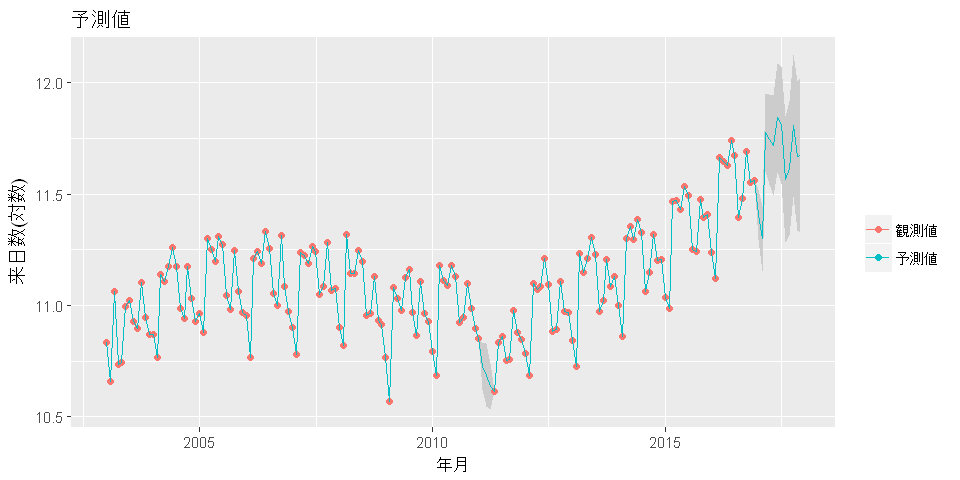
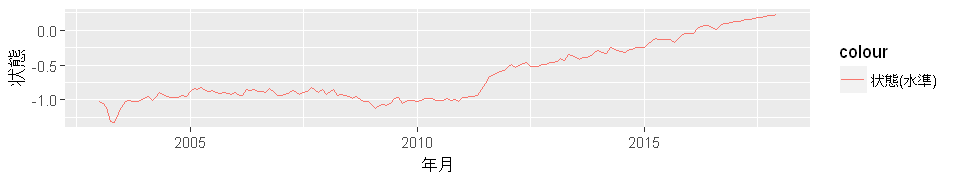
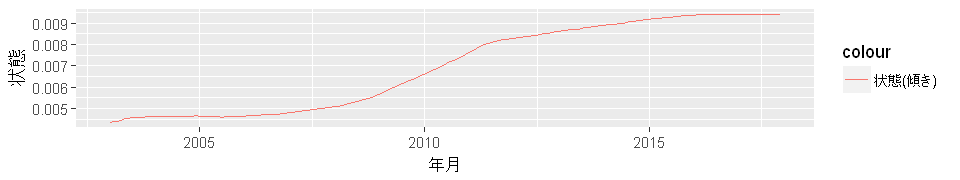


図9 予測結果(網掛けは予測値95％信頼区間) ※外れ値除去後





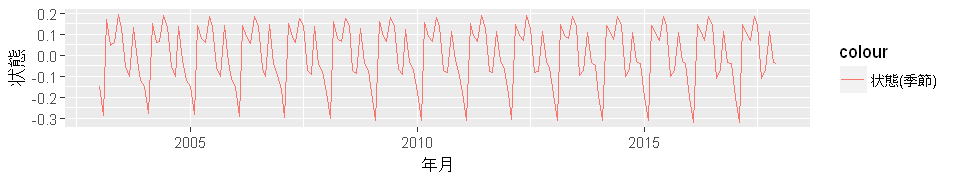


図10状態の推移 ※外れ値除去後

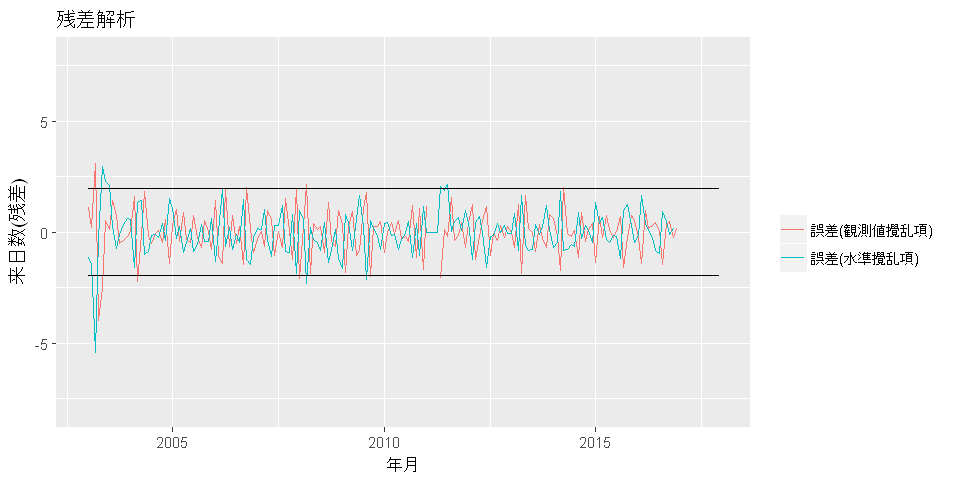


図11 残差 ※外れ値除去後

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。

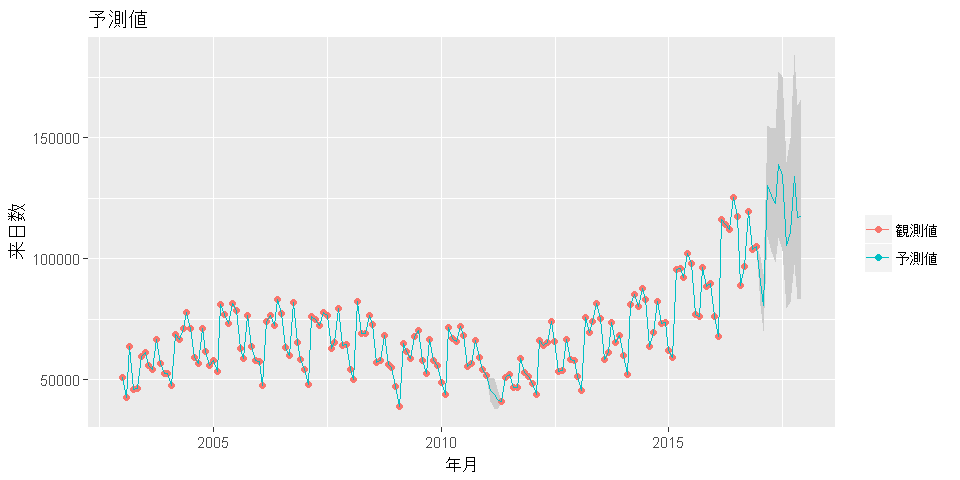


図12 予測結果(元のスケール)

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。

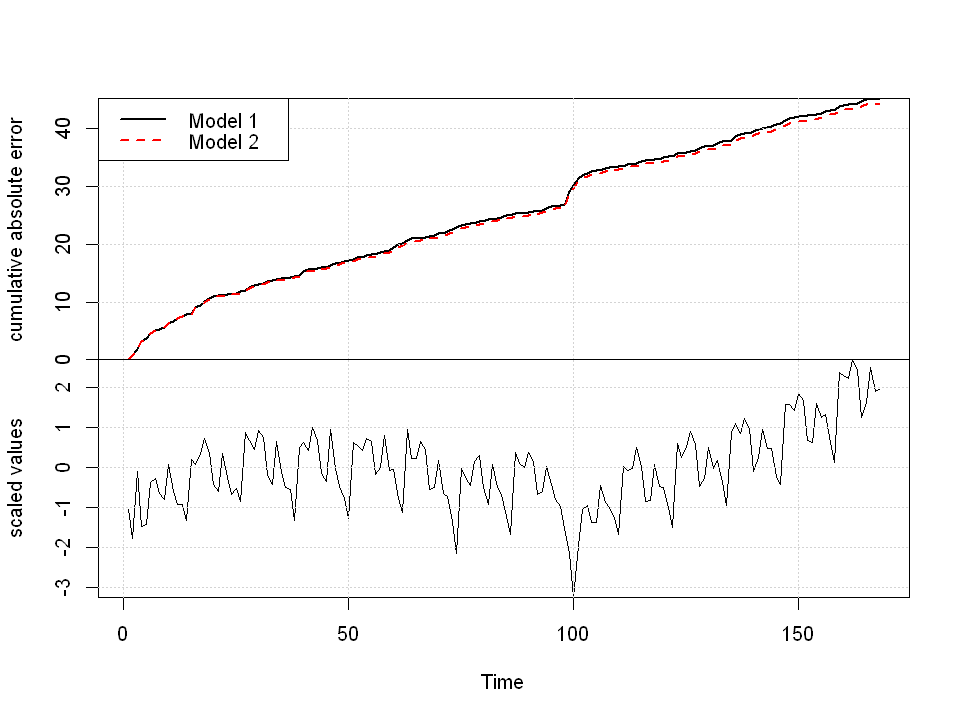


図13 累積誤差の比較 (横軸：データのインデックス)

　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。

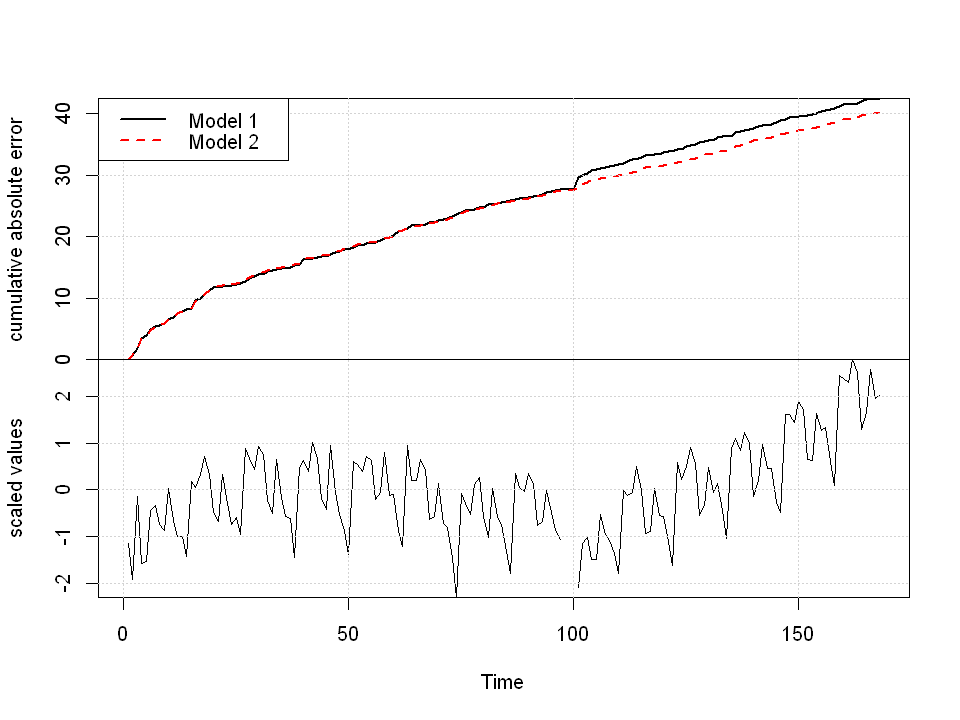


図14 累積誤差の比較 (横軸：データのインデックス)　※外れ値除去後

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，地震変数を加えたモデル(モデル2)の方がより良い精度であることがわかる。

　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。

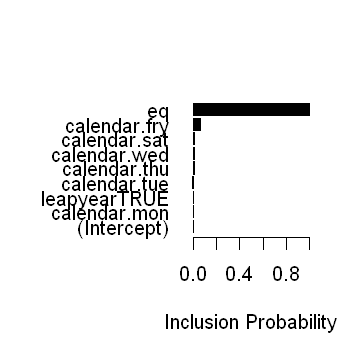


図15 モデル2の説明変数含有確率 (eq：地震，calendar.＊：カレンダー効果，leapyear：うるう年効果)

　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。