状態空間モデルの損害保険への活用
― Rパッケージ“KFAS”を用いた解析例―

野村　俊一

要旨

保険料率算定法は長年かけて発展を遂げており，例えばクレーム数のモデリングには，ポアソン分布などの離散分布が扱えてリスク区分ごとにクレーム頻度の推定が可能な一般化線形モデルを用いることができる。一方で，直近の保険金支払実績から料率算定を行っても，リスクは時間とともに変化するために将来のリスクを適切に予測できるとは限らない。時間的に変化するリスクを予測することは事業計画や料率算定などの経営戦略上重要であるが，トレンド変化を予測するのに線形回帰モデルでは不十分であり，適切な予測を行うには時系列モデルが必要となってくる。

そこで本稿では，状態空間モデルを用いたクレーム頻度の時系列モデリング法について紹介する。正規分布以外の確率分布を用いた線形状態空間モデルは，フリー統計解析ソフト”R”のパッケージ“KFAS”を利用することで手軽に実装することができる。ここでは，交通事故死傷者数の推移を解析する2つの解析例を，解析に用いたソースコードとともに紹介する。カウントデータのモデリングでしばしば問題となる過分散についても議論し，状態空間モデルへ過分散を導入する2つの方法を紹介する。

キーワード

クラス料率算定法，状態空間モデル，統計解析ソフトR，過分散，交通事故統計

1　はじめに

損害保険における純保険料率の設定には，期待保険金支払額に影響する契約属性情報（ファクター）の料率区分（クラス）ごとに料率が定まるクラス料率あるいはタリフ料率と呼ばれる形式が従来から用いられてきた。料率を定めるためのクラス料率算定法も古くから発展を遂げてきたが，近年では一般化線形モデルによる料率算定が主流となってきている。一般化線形モデルは，通常の最小二乗法による回帰分析とは異なり，正規分布以外の確率分布が扱えるため，カウントデータである事故件数や右に裾の長い分布をもつ保険金単価にも上手く適用することができる。また，tweedie 分布のように純保険料すなわち期待保険金総額を直接モデル化した確率分布を適用することもできる。

一般に料率算定において，料率の信頼性を保ちながら，できる限り最近のリスク水準を推定するために，過去どの程度の期間の事故データを利用するかがしばしば問題となる。事故件数は一部の未報告事故を除いて適時把握されるものの，特に保険金支払額は支払完了までに長期間を要するものもあるため，最近のデータだけに絞ると料率が過小推定される危険もある。さらに，自然災害や交通事故の傾向に見られるように，全体的な保険事故のリスクはある程度のトレンドをもちながら変化を続けており，直近のデータによく当てはまる料率を見積もれたとしても，それが将来においても適切な料率であるとは限らない。アンダーライターが契約ごとにリスクを評価して料率を定める企業向け保険商品は別として，個人向け保険商品では火災保険に限らず自動車保険なども近年は長期契約が増加しており，頻繁に料率改定を行ったとしても保有契約に新料率が反映されるまでには無視できないタイムラグがあるため，将来トレンドを見越して料率設定することで競合との競争を優位に進められる可能性がある。

リスクの時間変化のトレンドを見積もり予測するために，まず単純な方法として一般化線形モデルに時間項を設けて時間による変化率を推定することが考えられる。しかし，一般に線形回帰モデルは外挿に弱く，直近に変化したトレンドを繊細に捉えることができないため，適切な時系列モデルを導入して将来の時間変化を予測することが望ましい。料率算定に適した時系列モデルを構築するには様々な工夫が必要となるが，本稿では，柔軟な時系列モデルを構築でき，さらに簡易に解析することができる状態空間モデルを紹介する。

状態空間モデルとは，ある動的システムから観測される時系列の生成過程を，その生成過程を支配する状態と呼ばれる潜在変数の挙動と共に記述したモデルである。特に，潜在変数が状態方程式と呼ばれる線形式によって時間遷移し，観測値が正規分布に限らない一般の確率分布から生成される状態空間モデルは，線形非ガウス状態空間モデルと呼ばれ，一般化線形モデルの拡張形（動的一般化線形モデル）としても捉えられている。線形非ガウス状態空間モデルで推定を行うための効率的な解析手法は Durbin and Koopman [2000] により考案され，さらにプログラムとして実装するための C 言語ライブラリも開発された。最近では，フリー

の統計解析ソフトRにて線形非ガウス状態空間モデルを解析するためのパッケージ“KFAS”がHelske [2016]により開発され，短いコードで誰でも手軽に解析が行えるようになった。本稿は，線形非ガウス状態空間モデルの解析方法を簡略的に紹介するとともに，Rパッケージ“KFAS”を用いた解析の仕方について，交通事故データを用いた解析例とそのソースコードを詳しく解説することで，アクチュアリー実務への状態空間モデルの普及を目指すものである。

以降では，まず第2節にて線形非ガウス状態空間モデルを定義したのちに，第3節にてその解析手法の概要を紹介する。その後，第4節では交通事故による死者数の月次推移データ，第5節では交通事故による死傷者数の年次推移データについて，それぞれのモデル設計法とプログラム実装法を解説し，その解析結果を示していく。

2　線形非ガウス状態空間モデル

観測される時点をここでは離散時間として$t = 1, \ldots, n$とおき，各時点の観測値あるいは観測ベクトルを$y_1, \ldots, y_n$，状態ベクトルを$\alpha_1, \ldots, \alpha_n$と表す。状態空間モデルでは，観測される時系列の生成過程を観測モデルあるいは観測方程式と呼び，状態の動的システムをシステムモデルあるいは状態方程式と呼ぶ。

本稿では，係数行列を用いた線形方程式と，正規（ガウス）分布と非ガウスの分布からなる確率分布で定義される線形非ガウス状態空間モデルを扱う。まず観測モデルとして，時点$t = 1, \ldots, n$におけるy_tの確率（密度）関数を次のように与える。

$$p(y_t|y_1, \ldots, y_{t-1}, \alpha_1, \ldots, \alpha_t) = p(y_t|Z_t\alpha_t). \tag{1}$$

ここで，Z_tは時点tごとに定まる行列であり，右辺にある状態α_tの線形変換$\theta_t = Z_t\alpha_t$を信号と呼ぶ。この式は，過去時点の時系列と現在までの状態が与えられた下で，y_tは現在の状態の線形変換である信号$\theta_t = Z_t\alpha_t$にのみ依存することを示している。そして，各時点の観測モデルを支配する状態α_tの動的挙動は，時点$t = 2, \ldots, n$ごとに係数行列T_t, R_tが与えられた次式の状態方程式により記述されるものとする。

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t, \quad \eta_t \sim Normal(0, Q_t). \tag{2}$$

ここで，η_tは状態撹乱項と呼ばれる平均0，分散共分散行列Q_tの多変量正規分布に従う確率ベクトルであり，状態のランダムな挙動の源泉である。なお，初期時点の状態α_1の分布については，特定の確率分布あるいは定数が仮定されるか，あるいは定数κと単位行列Iを用いて$\alpha_1 \sim Normal(0, \kappa I)$と定義した上で，後に$\kappa \to \infty$の極限をとる散漫初期化と呼ばれる手法がとられる。通常，状態α_tの成分のうち状態方程式(2)が定常過程となるものについてはその定常分布が仮定され，ランダムウォークのような非定常過程となるものについては散漫初期化が行われる。

3　Rパッケージ“KFAS”による解析手法

前節にて紹介した式 (1)，(2) からなる線形非ガウス状態空間モデルは，観測モデルの一部の確率分布に対して，統計解析ソフトウェアRの外部パッケージ“KFAS”を利用して簡単に解析することができる。ここでは，Rパッケージ“KFAS”がとっているその解析手法の概要を紹介する。解析手法を全て解説するには紙面を取り過ぎるため，その詳細については野村 [2016] あるいは Durbin and Koopman [2012] を参照されたい。

上に述べたRパッケージ“KFAS”では，正規分布，ガンマ分布，二項分布，ポアソン分布，負の二項分布の5つの指数型分布族に属する確率分布を観測モデル (1) として利用できる。そのうち，本稿で主に用いるのは次の確率関数をもつポアソン分布である。

$$p(y_t|\theta_t) = \frac{(e^{\theta_t}u_t)^{y_t}}{y_t!}e^{-e^{\theta_t}u_t}, \quad y_t = 0, 1, \ldots. \tag{3}$$

ここで，u_t はエクスポージャを表す外部変数であり，このとき $\mathrm{E}(y_t|\theta_t) = \mathrm{Var}(y_t|\theta_t) = e^{\theta_t}u_t$ となる。さらに，観測値の分散がポアソン分布の分散を超える過分散の場合を扱うために，次の確率関数をもつ負の二項分布も用いる。

$$p(y_t|\theta_t) = \binom{u_t + y_t - 1}{y_t}\frac{e^{\theta_t y_t}u_t^{u_t}}{(e^{\theta_t} + u_t)^{u_t+y_t}}, \quad y_t = 0, 1, \ldots. \tag{4}$$

ここでの u_t は拡散パラメータと呼ばれ，$\mathrm{E}(y_t|\theta_t) = e^{\theta_t}$, $\mathrm{Var}(y_t|\theta_t) = e^{\theta_t} + e^{2\theta_t}/u_t$ となることから，負の二項分布の分散がポアソン分布の分散をどの程度超えるかに関わるパラメータであることがわかる。

Rパッケージ“KFAS”が採用する線形非ガウス状態空間モデルの解析手法は大きく3段階に分けられる。まず第1段階では，全時点の観測ベクトル $y = \{y_1, \ldots, y_n\}$ が与えられた下での全時点の信号 $\theta = \{\theta_1, \ldots, \theta_n\}$ の条件付き密度関数 $p(\theta|y)$ を最大化する条件付きモード $\hat{\theta} = \arg\max_\theta p(\theta|y)$ の値を求め，条件付きモードにおける正規近似により条件付き密度関数 $p(\theta|y)$ を近似する多変量正規分布 $g(\theta|y)$ を与える。ここで，条件付きモードの数値解は，カルマンフィルタとカルマンスムーザを含む効率的な逐次計算アルゴリズムにより得ることができる。続く第2段階では，全観測値 y が与えられた下での全時点の状態 $\alpha = \{\alpha_1, \ldots, \alpha_n\}$ のシミュレーション・サンプルを，真の条件付き分布 $p(\alpha|y)$ の代わりに前段階の近似分布を用いた多変量正規分布 $g(\alpha|y) = p(\alpha|\theta)g(\theta|y)$ よりサンプリングする。ここで，近似による分布差異の修正のために，得られた状態の各シミュレーション・サンプル $\alpha^{(i)}$, $i = 1, \ldots, N$ に対して，真の分布からのサンプル何個分に相当するかを表す重み w_i が真の分布と近似分布との密度関数比

$$w_i = \frac{p(y|\alpha^{(i)})}{g(y|\alpha^{(i)})} \tag{5}$$

によって与えられる。このように近似分布を用いて重み付きのシミュレーション・サンプルを得る手法はインポータンス・サンプリングと呼ばれている。最後の第3段階では，シミュレーション・サンプルから求めた標本平均，標本分散やパーセンタイルを用いて，モデルの状態推定や将来の観測値の予測，そして尤度の評価を行う。将来の観測値の予測では，将来時点 $t = n+1, n+2, \ldots$ の状態の各シミュレーション・サンプル $\alpha_t^{(i)}$ から観測値のサンプル $y_t^{(i)}$ を観測モデル $p(y_t|\alpha_t^{(i)})$ よりサンプリングして得られる。また，モデルの未知パラメータをまとめて ψ と表したときのモデルの尤度 $L(\psi)$ については，前段階の正規近似した分布における尤度 $L_g(\psi) = \int g(y|\alpha)p(\alpha)d\alpha$ が通常のカルマンフィルタによって求まるので，それと式 (5) の重みを用いて

$$\log \hat{L}(\psi) = \log L_g(\psi) + \log \frac{1}{N}\sum_{i=1}^{N} w_i \tag{6}$$

により近似的に得ることができる。さらに，式 (6) の尤度を用いてモデル選択のための赤池情報量規準（AIC: Akaike Information Criterion）が

$$\mathrm{AIC} = -2\log \hat{L}(\psi) + 2(r+q) \tag{7}$$

によって得られる。ただし，r はモデルの未知パラメータ ψ が取りうる値の空間の次元を表し，q は初期時点の状態 α_1 の成分のうち，散漫初期化を行った成分数を表す。

以上により，一部の確率分布を用いた線形非ガウス状態空間モデルから尤度を用いたパラメータ推定および各時点の状態推定（状態平滑化）および観測値の予測が行われる。R パッケージ“ KFAS ”を用いると，上記のプロセスが関数の内部処理で適切に実行されるため，解析手法の仔細まで理解せずとも線形非ガウス状態空間モデルによる高度な解析を行うことができる。次節以降の交通事故データの解析例では，実際の時系列データに対する状態空間モデルの設計方法に加え，R パッケージ“ KFAS ”を用いて解析を実行するためのコーディング例を紹介していく。

4　交通事故による月別死者数の解析例

本節では，警察庁の公表する『交通事故発生状況』の統計表より取得した，交通事故による月別死者数の時系列データに対して線形非ガウス状態空間モデルを適用した解析例を紹介する。

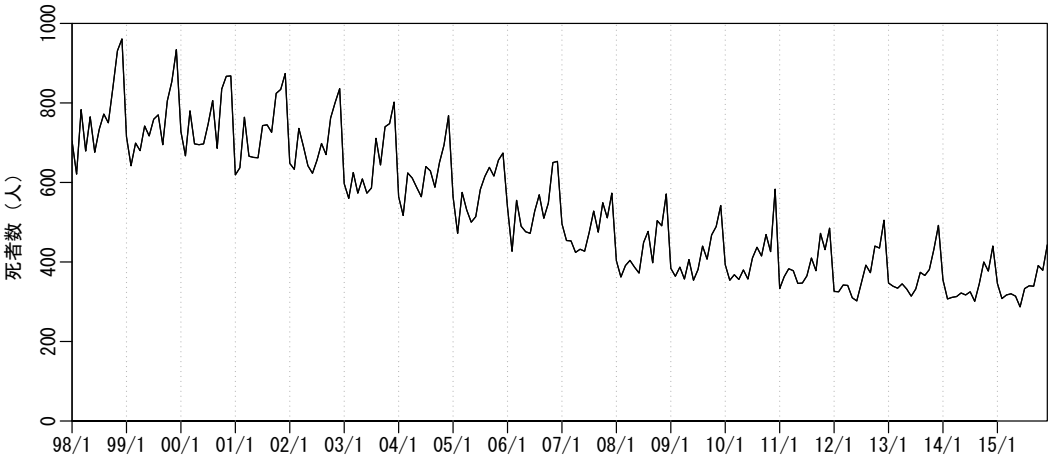

図 1　交通事故による月別死者数の推移

4.1　モデルの定式化

まず，図 1 に示された 1998 年 1 月から 2015 年 12 月までの交通事故による月別死者数に対して，線形非ガウス状態空間モデルを用いて解析を行う。図の推移から読み取れることとして，月別死者数には明らかな年周期の季節変動が存在しており，また緩やかな減少トレンドが続いていることがわかる。そこで，各時点 $t = 1, \ldots, 216$ の月別死者数の観測値 y_t に対する観測モデルと，その背景にある状態方程式を次式のように定義したモデルを考える。

$$y_t \sim Poisson(e^{\theta_t}), \tag{8}$$

$$\theta_t = \mu_t + \gamma_t, \tag{9}$$

$$\mu_t = \mu_{t-1} + \nu_{t-1}, \tag{10}$$

$$\nu_t = \nu_{t-1} + \eta_{t1}, \quad \eta_{t1} \sim Normal(0, \sigma_1^2), \tag{11}$$

$$\gamma_t = -(\gamma_{t-1} + \cdots + \gamma_{t-11}) + \eta_{t2}, \quad \eta_{t2} \sim Normal(0, \sigma_2^2) \tag{12}$$

ここで，式 (8) は観測値 y_t が式 (3) において $u_t = 1$ としたポアソン分布に従うことを表している。エクスポージャ u_t には本来，各時点における全国の人口などを設定すべきであるが，1998 年以降の全国の人口変化は 2 百万人未満であり大きな変化はないことから，簡便のためエクスポージャ一定としている。

次の式 (9) では，観測値 y_t の期待値の対数である信号 θ_t が，2 つの状態 μ_t と γ_t の和として定義されている。状態成分 μ_t の挙動を示す式 (10) と (11) は，2 次のトレンド成分モデルあるいは平滑化トレンドモデルと呼ばれ，それに従う μ_t は水準成分と呼ばれる。水準成分 μ_t の各時点における変化量は別の状態成分 ν_t により表されており，この ν_t は傾き成分と呼ばれる。水準成分の傾き ν_t が撹乱項 η_{t1} により徐々に変化することにより，水準成分はトレンドをもって滑らかに変化することとなる。

また，他方の状態成分 γ_t が従う式 (12) は季節成分モデルと呼ばれ，それに従う γ_t は季節成分と呼ばれる。式 (12) において撹乱項 η_{t2} を省き，さらに $\sum_{t=1}^{12}\gamma_t = 0$ という制約を設けると，$\gamma_{13} = \gamma_1, \gamma_{14} = \gamma_2, \cdots$ というように季節成分 γ_t は 12ヵ月周期で同じ値を繰り返すことが確かめられる。これに撹乱項 η_{t2} が加わることで，季節変動のしかたが徐々に変化することを許したのが式 (12) となっている。

ここまでの式 (10)，(11)，(12) はいずれも状態成分の動的挙動を表す状態方程式の一部であり，すべてをまとめて状態ベクトルを $\alpha_t = (\mu_t\ \nu_t\ \gamma_t\ \cdots\ \gamma_{t-9}\ \gamma_{t-10})'$（$'$ はベクトルまたは行列の転置を表す）と定義した次の状態方程式 (2) に対応した形へと書き直すことができる。

$$\begin{pmatrix} \mu_t \\ \nu_t \\ \gamma_t \\ \vdots \\ \gamma_{t-9} \\ \gamma_{t-10} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & -1 & \cdots & -1 & -1 \\ \vdots & \vdots & 1 & & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_{t-1} \\ \nu_{t-1} \\ \gamma_{t-1} \\ \vdots \\ \gamma_{t-10} \\ \gamma_{t-11} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{t1} \\ \eta_{t2} \end{pmatrix}, \tag{13}$$

$$\begin{pmatrix} \eta_{t1} \\ \eta_{t2} \end{pmatrix} \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right) \tag{14}$$

ここで，式 (14) では撹乱項 η_{t1} と η_{t2} が互いに無相関（正規分布なのですなわち独立）であると仮定しており，それぞれの分散 σ_1^2, σ_2^2 は推定すべき未知パラメータとして扱う。

4.2 R パッケージ“KFAS”による解析

以上では式 (8) から (12) により定式化したモデルが線形非ガウス状態空間モデル (1)，(2) として記述できることが確かめられた。次に，そのモデルを R パッケージ“KFAS”を用いて定義し，パラメータ推定，状態平滑化および対数尤度と AIC の算出を行うための R ソースコード例を以下に示す。

```
# パッケージ“KFAS”の読み込み
library(KFAS)
# 状態空間モデルの定義（ポアソン分布）
modPois <- SSModel(tsuki ~ SSMtrend(2, Q=list(0,NA)) + SSMseasonal(12, Q=NA),
  distribution="poisson", u=1)
# 対数尤度最大化による未知パラメータの推定
fitPois <- fitSSM(modPois, c(-15,-10), nsim=1000, method="BFGS")
# AIC の算出
aicPois <- 2 * fitPois$optim.out$value + 2 * (2 + 13)
# インポータンス・サンプリングによる状態推定
kfsPois <- KFS(fitPois$model, nsim=1000)
```

最初のコードは，状態空間モデルに関する様々な関数が収録されたRパッケージ“KFAS”を読み込んでいる。パッケージ“KFAS”はRにプリインストールされてはいないため，事前にinstall.packages(“KFAS”)と実行してミラーサーバからのダウンロードとインストールを済ませておく必要がある。

パッケージを読み込んだら，まずは関数SSModelを用いてモデルの定義を行う。SSModelの最初の引数では，チルダ~の左側に月別死者数を格納した時系列データtsukiを与え，右側にモデル式を与えている。チルダ~の右側では，2つの関数SSMtrendとSSMseasonalが足されており，前者はトレンド成分モデルに従う水準成分μ_tを，後者は季節成分モデルに従う季節成分γ_tを定義する関数となっている。

SSMtrendの最初の引数にはトレンド成分モデルの次数2を定め，次の引数Qには撹乱項の分散をリスト形式で定めている。リストの第1要素は，水準成分に直接加わる撹乱項の分散であり，ここでは式(10)に従い分散0すなわち撹乱項なしとしておく。リストの第2要素は，式(11)で傾き成分に加わる撹乱項η_{t1}の分散σ_1^2を指しており，後に推定するためにここでは未知パラメータであることを示すNAを与えておく。続いて，SSMseasonalの最初の引数には季節成分の周期12を定め，次の引数QにはSSMtrendと同様に撹乱項の分散を未知パラメータNAとして定義している。

関数SSModelの次の引数distributionは，観測値が従う確率分布を指定するものであり，ここではポアソン分布"poisson"を指定している。最後の引数uはポアソン分布(3)のエクスポージャu_tを定める部分であり，上述のとおり一律$u_t = 1$を設定しておく。

以上により定義されたモデルが，変数modPoisに格納されている。変数modPoisの収録内容は例えばmodPois[T]などと実行すれば一挙に表示できる。例えば，$Tの後に続いて表示される係数行列$T_t$について，式(13)に示した係数行列と一致することが確かめられる。また，$Qの後に表示される撹乱項分散行列$Q_t$は対角成分にNAが2つ並んでおり，これらが次に推定される未知パラメータσ_1^2, σ_2^2に該当する。

次のコードにある関数fitSSMは，モデルmodPoisの未知パラメータを負の対数尤度の最適化（最尤法）により推定する関数である。1番目の引数にはモデルが格

納された変数 modPois を与え，2 番目には未知パラメータの初期値を対数で指定している。ここで，モデルの負の対数尤度は必ずしも凸関数にはならず，初期値に依存して局所最適解に収束することが多々あるため，様々な初期値のパターンを用意して大域的な最適解を探索する必要があることに注意する。そして，関数 fitSSM の 3 番目の引数 nsim では対数尤度を式 (6) で近似的に求めるためのシミュレーション・サンプルのサイズ $N = 1000$ を指定し，4 番目の引数 method では最適化手法として疑似ニュートン法"BFGS"を指定している。なお，実際の最適化計算は R の最適化関数 optim に引き渡されているため，関数 optim のオプション指定のための引数はすべて使用可能である。関数 fitSSM の実行結果が入った変数 fitPois には，関数 optim による最適化の返り値が fitPois$optim.out に，モデル modPois の未知パラメータ NA が最尤推定値へと置き換えられたものが fitPois$model に格納されている。そのため，fitPois$optim.out$value により，最適化された最尤推定値に対する負の対数尤度を得ることができる。

続くコードでは，先ほど最適化された負の対数尤度 fitPois$optim.out$value を用いて AIC を計算している。ここで，式 (7) における未知パラメータ σ_1^2, σ_2^2 の次元は $r = 2$ であり，散漫初期化された状態成分数は $q = 13$ である。状態方程式 (10)，(11)，(12) からわかるように各状態成分はランダムウォークのような非定常な挙動をするため，関数 SSMtrend と SSMseasonal が定義するトレンド成分モデルと季節成分モデルはデフォルトで全ての状態成分に対して散漫初期化が行われる。

最後の関数 KFS では，最初の引数のモデル fitPois$model に対してインポータンス・サンプリングを実行して引数 nsim で指定された数 $N = 1000$ のシミュレーション・サンプル $\alpha^{(1)}, \ldots, \alpha^{(N)}$ を取得し，各サンプルの重み w_i に基づく加重平均 $\hat{\alpha} = \sum_{i=1}^{N} w_i \alpha^{(i)} / \sum_{i=1}^{N} w_i$ により状態の推定を行っている。関数 KFS の実行結果が入った変数 kfsPois には，各時点の状態 α_t の推定値とその標準誤差が kfsPois$alphahat および kfsPois$V に，各時点の信号 $\theta_t = Z_t \alpha_t$ に基づく観測値の期待値 $\mathrm{E}(y_t|\alpha_t) = e^{\theta_t}$ の推定値とその標準誤差が kfsPois$muhat および kfsPois$V_mu にそれぞれ格納されている。

以上で R パッケージ“KFAS”を用いた状態空間モデルの一連の解析方法を示した。ここで，kfsPois に格納された各時点の状態成分の推定値 kfsPois$alphahat および観測値の期待値の推定値 kfsPois$muhat の推移を図 2 と図 3 にそれぞれ示した。図 2(a) の水準成分は観測期間中ずっと減少を続けているものの，図 2(b) に示されたその傾きについては増減を繰り返していることがわかる。また図 2(c) からは，季節成分による季節変動が毎年一定ではなく，徐々に変化している様子が見てとれる。図 3 では，赤線で示した期待値の推定値が概ね観測値に近い値をとっており，大きく乖離した外れ値も見られない。

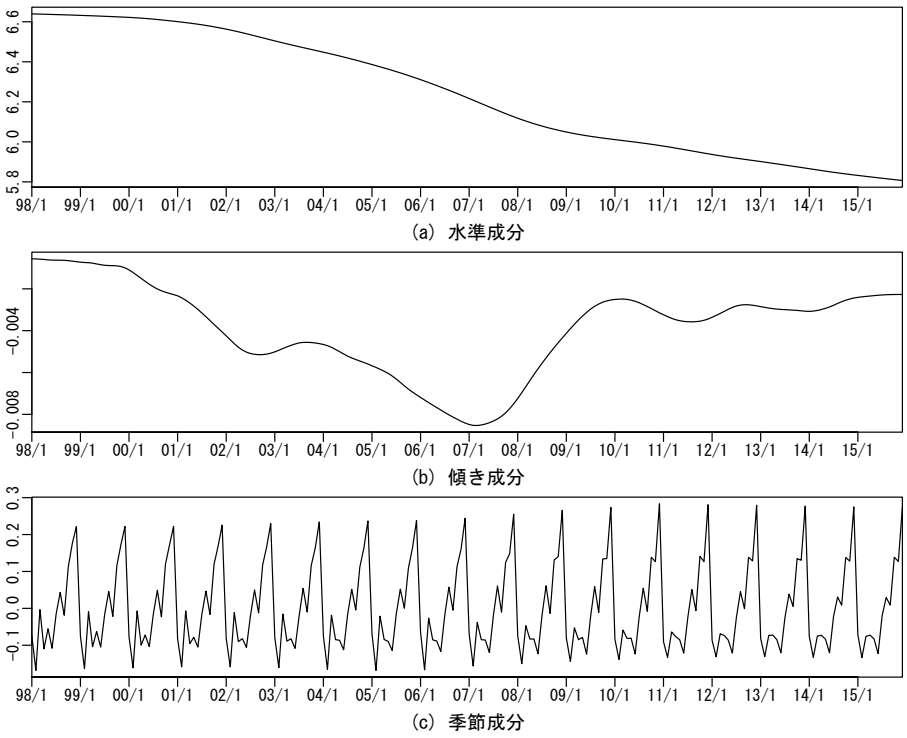

図 2　各状態成分の推定値の推移

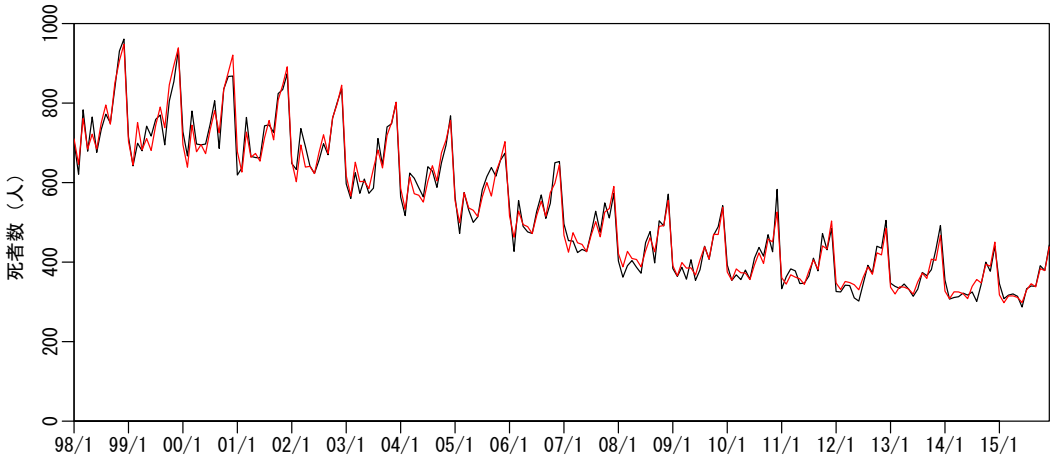

図 3　月別死者数の期待値の推定値の推移

4.3　負の二項分布モデルによる過分散の検討

ここで，観測モデルの分散がポアソン分布よりも大きい過分散である可能性を検討するため，式 (8) に代えて次式のように観測モデルに負の二項分布を仮定したモデルを代替候補として用意する。

$$y_t \sim NegativeBinomial(e^{\theta_t}, u). \tag{15}$$

ただし，右辺の u は式 (4) における拡散パラメータを指す。この負の二項分布を用いた状態空間モデルによる解析コードを以下に示す。

```
# 状態空間モデルの定義（負の二項分布）
modNeg <- SSModel(tsuki ~ SSMtrend(2, Q=list(0,NA)) + SSMseasonal(12, Q=NA),
  distribution="negative binomial", u=1)
# モデル更新関数（未知パラメータの値を更新するための関数）の作成
updatefn <- function(pars, model){
  model$Q[2,2,] = exp(pars[1])
  model$Q[3,3,] = exp(pars[2])
  model$u[,] = exp(pars[3])
  return(model)
}
# 対数尤度最大化による未知パラメータの推定
fitNeg <- fitSSM(modNeg, c(-15,-10,8), updatefn, method="BFGS", nsim=1000)
# AIC の算出
aicNeg <- 2 * fitNeg$optim.out$value + 2 * (3 + 13)
# インポータンス・サンプリングによる状態推定
kfsNeg <- KFS(fitNeg$model, nsim=1000)
```

上記のコードは，前に示したポアソン分布のものと基本的には同じであり，関数 SSModel によるモデル定義は引数 distribution をポアソン分布"poisson"から負の二項分布"negative binomial"に変更した以外に差異はない。

ポアソン分布との大きな違いは，未知パラメータとして撹乱項分散 σ_1^2, σ_2^2 の他に拡散パラメータ u も推定する点にある。撹乱項分散とは異なり，拡散パラメータは NA としても関数 fitSSM の推定対象にならず，そのため関数 fitSSM の引数にモデル内のパラメータ更新を行う関数 updatefn を追加している。関数 updatefn は上のコードのようにユーザーで定義しなければならず，パラメータ更新用の数値 pars と更新されるモデル model を引数として用意し，撹乱項分散 model$Q と拡散パラメータ model$u が更新されたモデル model を返り値とする。なお，パラメータ更新の際に引数 pars の指数をとっているのは，更新するパラメータ値は正値に限定されるのに対し，引数 pars の定義域を実数全体にしておくための措置である。

その後は，AIC の算出においては未知パラメータの次元が 2 から 3 に増えたこと以外に差異はなく，関数 KFS による状態推定においてはポアソン分布のときと全く違いはない。

解析の結果，ポアソン分布のモデルで算出された AIC は 2147.8，負の二項分布のモデルで算出された AIC は 2149.2 となり，本解析においてはポアソン分布の方が適切であり，過分散の兆候は見られないものと結論付けられる。

4.4 月別死者数の将来予測

ここまでの解析では尤度によるパラメータ推定，AIC の算出，各時点の状態推定までを行ってきたが，最後に将来の観測値を予測するための方法を示す。以下

のコードでは，直近 12ヵ月間（2015 年 1〜12 月）の観測値を敢えて欠測値 NA に代えることで，2015 年中の未知の観測値を予測させて，実際の観測値との答え合わせができる状況を作っている。

```
# 直近 12ヵ月間（2015 年 1〜12 月）のデータを欠測値（NA）に代えたデータを作成
tsukiNA <- tsuki; tsukiNA[205:216] <- NA
# 状態空間モデルの定義（ポアソン分布）
modPoisNA <- SSModel(tsukiNA ~
  SSMtrend(2, Q=list(0,NA)) + SSMseasonal(12, Q=NA),
  distribution="poisson", u=1)
# 対数尤度最大化による未知パラメータの推定
fitPoisNA <- fitSSM(modPoisNA, c(-15,-10), method="BFGS", nsim=1000)
# インポータンス・サンプリングによる状態推定
kfsPoisNA <- KFS(fitPoisNA$model, nsim=1000)
# インポータンス・サンプリングによる観測値の予測
prePoisNA <- predict(fitPoisNA$model, interval="prediction",
  level=0.95, nsim=10000)
```

上記では，状態推定を行うまでデータを改変した以外の変更点はない。最後の観測値の予測には，パッケージ“KFAS”に内蔵された関数 predict を利用している。関数 predict の最初の引数にはモデル fitPoisNA$model を渡し，次の引数 interval は"prediction"と指定することで観測値の予測値と予測区間の下限および上限が返されるようになる。なお，引数 interval に代わりに"confidence"と指定すると，状態の推定値と信頼区間を得ることもできる。その後は，引数 level にて予測区間の確率を 0.95 すなわち 95%予測区間と指定し，引数 nsim にてシミュレーション・サンプルのサイズを $N = 10000$ と精度向上のため大きめにとっている。

以上のコードによって得られた 2015 年 1〜12 月の観測値に対する予測値および 95%予測区間を，実際の観測値と共に図 4 に示した。さらに，直近 6ヵ月間の観測値のみを欠測値に代えて同様に予測した結果も図 4 に示している。いずれの予測においても，実際の観測値は予測区間内に収まっており概ね良い予測を与えている。しかし，赤線で示された直近 12ヵ月間で予測した予測値は，全体的に実際の観測値より下ぶれしていることが見てとれ，2015 年合計では約 200 人過小に予測されたことになる。これは，予測期間直前の 2013 年から 2014 年では大きく減少傾向にあったために，状態成分のうち増減トレンドを表す傾き成分が低く推定されて 2015 年もそのままの減少傾向で予測されたのに対して，実際には 2015 年は 2014 年と概ね横ばいに推移したことが原因となっている。このように長期的なトレンドの予測には上ぶれ・下ぶれのリスクを伴うが，一方で青線で示された直近 6ヵ月間で予測した予測値には大きな下ぶれは見られず，2015 年上半期の傾向から傾き成分の推定が上方修正されたことを示している。

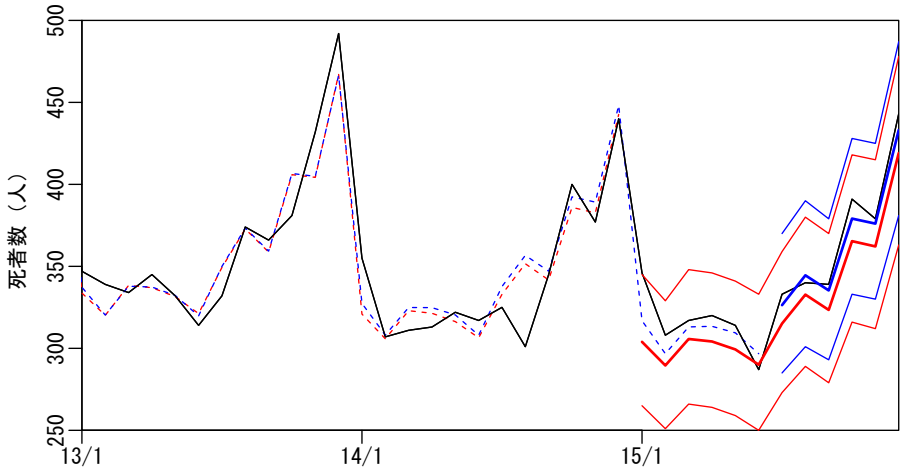

図 4　直近 12ヵ月間の予測（赤線）と直近 6ヵ月間の予測（青線）の比較

以上のような月別推移データの解析によるトレンド予測は，例えば会社の業績予想について，年度初めに予想していたトレンドを年度途中に修正すべきか検討するようなシチュエーションで，勘に頼らない統計的な評価を与えてくれる有効な手段と言えよう。

5　交通事故による年間死傷者数の解析例

前節では交通事故による死者数の月次推移を解析したが，本節では同じく警察庁の『交通事故発生状況』の統計表より，集計期間を年単位にする代わりに，死者数に加えて重傷者数およびその他の負傷者数（負傷者数 – 重傷者数）の 3 つの観測値が推移する多変量時系列を解析する。

図 5 には 1980 年から 2015 年までの各年の人口 10 万人あたり死者数，重傷者数およびその他の負傷者数の推移を示している。図 5 からは，3 者の推移が概ね近い傾向をとっていることが見てとれ，特に死者数と重傷者数の推移がとても似通っていることがわかる。そのため，多変量時系列としてモデル化する際には，3 つの観測値間に存在する相関関係を考慮に入れることが予測にとって重要となる。

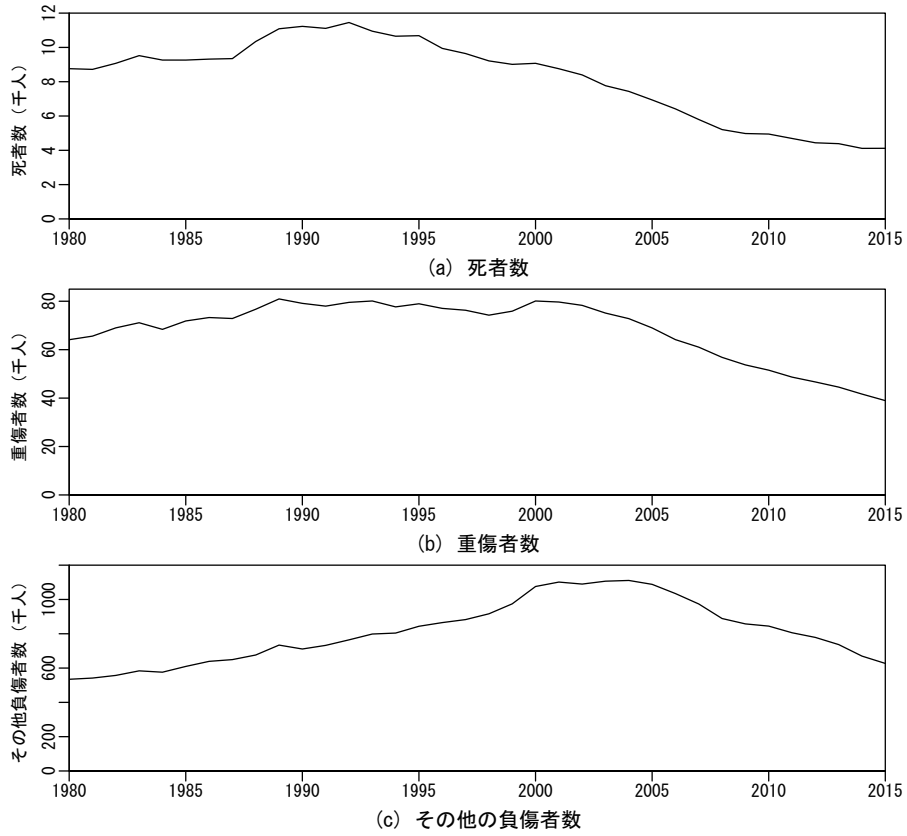

図 5　交通事故による年間死傷者数の推移

5.1　モデルの定式化

各時点 $t = 1, \ldots, n$ における年間死者数，重傷者数，その他の負傷者数の観測値を $y_t = (y_{t1}\ y_{t2}\ y_{t3})'$ とベクトルで表す。多変量時系列をモデル化する場合，まずは個別の単変量時系列としてモデル設計するところから始めるのがよい。各成分 $i = 1, 2, 3$ に対して，次のようなモデルを考える。

$$y_{ti} \sim Poisson(u_t e^{\theta_{ti}}), \tag{16}$$

$$\theta_{ti} = \mu_{ti}, \tag{17}$$

$$\mu_{ti} = \mu_{t-1,i} + \nu_{t-1,i}, \tag{18}$$

$$\nu_{ti} = \nu_{t-1,i} + \eta_{ti}, \quad \eta_{ti} \sim Normal(0, \sigma_i^2) \tag{19}$$

式 (16) は観測値の各成分 y_{t1}, y_{t2}, y_{t3} がそれぞれ式 (3) のポアソン分布に従うことを表しており，ここでのエクスポージャ u_t は総務省統計局『10 月 1 日現在推計人口』および『国勢調査結果』による各年 10 月 1 日現在における全国の推計人口である。図 6 に示したように，1980 年から 2015 年までの間に 1 千万人程度人口が伸びていることがわかる。

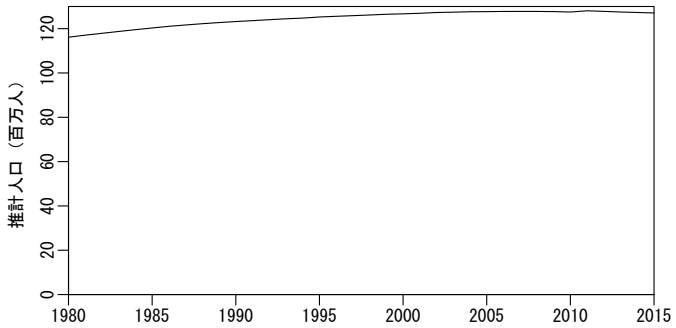

図 6　全国の推計人口の推移

式 (18) と (19) は前節と同じ 2 次のトレンド成分モデルを表している。前節の月次データとは異なり，年次データであるため年内の季節変動を考慮する必要はなく，そのため式 (17) において前節のような季節成分は外れている。

式 (16) から (19) までは成分ごとに独立してモデル化されており，ここまでは単変量時系列として扱うのと何ら変わらず多変量時系列にする意味はない。しかし，次のように撹乱項 $\eta_{t1}, \eta_{t2}, \eta_{t3}$ の成分間に相関をもたせることで，状態および観測ベクトルの各成分が連動して推移するため，多変量時系列として扱うメリットが生まれることとなる。

$$\begin{pmatrix} \eta_{t1} \\ \eta_{t2} \\ \eta_{t3} \end{pmatrix} \sim Normal\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix} \right). \tag{20}$$

このようなモデルは，式 (16) から (19) までの時系列方程式には成分間の関係が表れないにも関わらず，状態の撹乱項を通じて成分間の相関関係が生ずることから，一見無関係な時系列方程式モデル（SUTSE: seemingly unrelated time seris equations model）と呼ばれる。

以上のモデルも，前節と同様に状態を $\alpha_t = (\mu_{t1}\ \nu_{t1}\ \mu_{t2}\ \nu_{t2}\ \mu_{t3}\ \nu_{t3})'$ とおいて，ひとつの状態方程式で次のように表現することができる。

$$\begin{pmatrix} \mu_{t1} \\ \nu_{t1} \\ \mu_{t2} \\ \nu_{t2} \\ \mu_{t3} \\ \nu_{t3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{t-1,1} \\ \nu_{t-1,1} \\ \mu_{t-1,2} \\ \nu_{t-1,2} \\ \mu_{t-1,3} \\ \nu_{t-1,3} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{t1} \\ \eta_{t2} \\ \eta_{t3} \end{pmatrix}, \tag{21}$$

このとき，観測モデル (16) における信号 $\theta_t = (\theta_{t1}\ \theta_{t2}\ \theta_{t3})' = Z_t\alpha_t$ は

$$\begin{pmatrix} \theta_{t1} \\ \theta_{t2} \\ \theta_{t3} \end{pmatrix} = \begin{pmatrix} \mu_{t1} \\ \mu_{t2} \\ \mu_{t3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \alpha_t \tag{22}$$

と表されることとなる。

5.2 Rパッケージ“KFAS”による解析

以上の式 (16)，(20)，(21)，(22) からなる線形非ガウス状態空間モデルの解析コード例は下記のようになる。

```
# 直近 5 年間（2011〜2015 年）のデータを欠測値（NA）に代えたデータを作成
nenNA <- nen; nenNA[32:36,1:3] <- NA
# 状態空間モデルの定義
modPois2 <- SSModel(nenNA[,1:3] ~
  SSMtrend(2, Q = list(matrix(0, 3, 3), matrix(NA, 3, 3))),
  distribution = "poisson", u = nenNA[,4])
# 対数尤度最大化による未知パラメータの推定
fitPois2 <- fitSSM(modPois2, c(rep(-10,3),rep(0,3)), method="BFGS",nsim=1000)
# インポータンス・サンプリングによる状態推定
kfsPois2 <- KFS(fitPois2$model, nsim=1000)
# AIC の算出
aicPois2 <- 2 * fitPois2$optim.out$value + 2 * (6 + 6)
# インポータンス・サンプリングによる観測値の予測
prePois2 <- predict(fitPois2$model, interval="prediction",
  level=0.95, nsim=10000)
```

解析データ nen の各列には年ごとの死者数，重傷者数，その他の負傷者数（負傷者数 – 重傷者数），推計人口が入っている。ここでは予測性能を評価するために 2011〜2015 年の直近 5 年間の死傷者数を欠測値に代えたデータ nenNA を用いる。

関数 SSModel による状態空間モデルは，多変量時系列に対しても上記のように単純なコードで定義することができる。最初の引数では，チルダ~の左側に年間死傷者数のデータが死者数，重傷者数，その他の負傷者数の 3 列分与えられている。チルダ~の右側では関数 SSMtrend を用いて 2 次のトレンド成分モデルを定義しているが，多変量時系列になって変わったのは撹乱項分散行列 Q を与えるリストの各要素が行列となった点のみである。前節と同様の考え方により，リストの第 1 要素は全成分 0 のゼロ行列，第 2 要素は全成分を未知パラメータ NA とした行列としている。最後の引数 u ではデータ nen の 4 列目に格納された各年の推計人口をエクスポージャとして与えている。

続く関数 fitSSM による未知パラメータの推定では，初期値として 6 つの値を指定している。ここでの未知パラメータは式 (20) の撹乱項分散行列であり，そのパラメータ数は 6 であることがわかる。ただし，fitSSM が扱う撹乱項分散行列のパラメトリゼーションは式 (20) と異なり，パラメータの定義域が実数全体をとるよう工夫されている。

パラメータ推定以降は前節と同様のコードで解析できる。AIC の算出にあたっては，未知パラメータの次元は 6 で，状態の全 6 成分ともトレンド成分モデルのデフォルトで散漫初期化されているため，式 (7) において $r = q = 6$ となる。

以上の解析結果に基づいて算出された，各年の 10 万人あたり死傷者数の期待値 $e^{\theta_{ti}}$ の推定値と，将来の予測値および予測区間について，図 7 にそれぞれ赤線と緑線で示した。直近 5 年間の観測値はいずれも 95%予測区間の範囲内に収まってい

るが，年が進むにつれ予測区間の幅と予測値からの乖離が大きくなっており，長期の予測は予測精度を大きく落とすことがわかる。

また，図 7 の青線と水色線は，式 (20) の分散共分散行列を零行列（全ての成分を零）とした場合，すなわち，実質的にポアソン分布と対数リンク関数による一般化線形モデルを適用した場合の，死傷者数の期待値 $e^{\theta_{ti}}$ の推定値および将来の予測値を示している。このポアソン回帰の結果はいずれもデータから大きく乖離しており，特に，その他負傷者数の将来予測は誤った増加トレンドに従って非常に的外れな予測をしている。このように，時系列に対して時間を説明変数とした回帰を行うことは，しばしば非常に誤った予測を与えることとなるため，時系列の予測（外挿）問題には状態空間モデルのような時系列モデルを適用すべきである。

なお，予測前の期間では，赤線の期待値の推定値が黒線の観測値とほぼ重なっており，モデルによる状態推定が観測値にオーバーフィッティングしている可能性が疑われる。実際，水準成分がトレンドを保って滑らかに推移するよう 2 次のトレンド成分モデルを採用しているにも関わらず，赤線の期待値は観測値に合わせて傾きの符号を何度も反転させていることも，オーバーフィッティングの兆候といえる。これは，観測モデルであるポアソン分布の標準偏差が分散 = 期待値の平方根となり非常に小さいことが主原因と考えられ，次に紹介するように過分散のモデルに替えることで解決することができる。

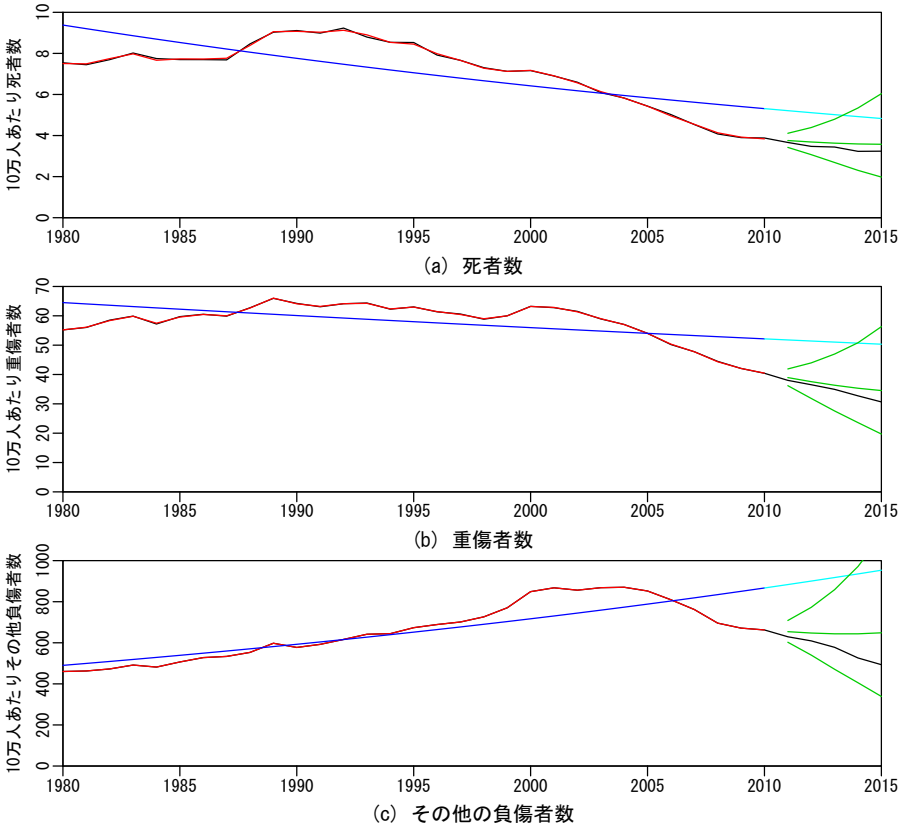

図 7　直近 5 年間の年間死傷者数の予測

5.3 変量効果の導入による過分散の検討

前節の月別死者数の解析では，過分散のモデルとして負の二項分布を取り入れたが，ここでは代わりに変量効果の導入により過分散を表現した次のモデルを観測ベクトルの各成分 $i = 1, 2, 3$ に対して導入する。

$$y_{ti} \sim Poisson(u_t e^{\theta_{ti}}), \tag{23}$$

$$\theta_{ti} = \mu_{ti} + \xi_{ti}, \tag{24}$$

$$\mu_{ti} = \mu_{t-1,i} + \nu_{t-1,i}, \tag{25}$$

$$\nu_{ti} = \nu_{t-1,i} + \eta_{ti}, \quad \eta_{ti1} \sim Normal(0, \sigma_{i1}^2), \tag{26}$$

$$\xi_{ti} = \eta_{ti2}, \quad \eta_{ti2} \sim Normal(0, \sigma_{i2}^2) \tag{27}$$

式 (24) にて信号 θ_{ti} に加えられた項 ξ_{ti} が変量効果と呼ばれるものである。式 (27) からわかるように変量効果 $\xi_{t1}, \xi_{t2}, \xi_{t3}$ は過去の値とは独立なその年特有のノイズとして与えられており，ポアソン分布の期待値 $e^{\theta_{ti}}$ に揺らぎを加えることで観測値の分散を増幅させる効果を持っている。ここでは，トレンド成分モデルの撹乱項 $\eta_{t1} = (\eta_{t11}\ \eta_{t21}\ \eta_{t31})'$ と変量効果の撹乱項 $\eta_{t2} = (\eta_{t12}\ \eta_{t22}\ \eta_{t32})'$ が互いに独立に式 (20) のような成分間の相関をもつ多変量正規分布に従うことを仮定して，相関のある多変量時系列モデルとして解析を行う。

上記のモデルの解析コードを以下に示す。

```
# 状態空間モデルの定義
modPois3 <- SSModel(nenNA[,1:3] ~
  SSMtrend(2, Q = list(matrix(0, 3, 3), matrix(NA, 3, 3))) +
  SSMcustom(Z=diag(3), T=matrix(0, 3, 3), Q=matrix(NA, 3, 3)),
  distribution = "poisson", u = nenNA[,4])
# 対数尤度最大化による未知パラメータの推定
fitPois3 <- fitSSM(modPois3, c(rep(-10,6),rep(0,6)), method="BFGS",nsim=1000)
# AIC の算出
aicPois3 <- 2 * fitPois3$optim.out$value + 2 * (12 + 6)
# インポータンス・サンプリングによる観測値の予測
prePois3 <- predict(fitPois3$model, interval="prediction",
  level=0.95, nsim=10000)
```

変量効果の定義には，任意の状態方程式を定義できる関数 SSMcustom を用いている。式 (24) と (27) により変量効果 $\xi_{t1}, \xi_{t2}, \xi_{t3}$ の部分に関する係数行列は $Z_t = I_3$（3 次元単位行列），$T_t = O$（ゼロ行列）となり，撹乱項分散行列はトレンド成分モデルの撹乱項と同様に全成分 NA（未知パラメータ）としておく。よって，未知パラメータは先ほどの倍の 12 に増え，関数 fitSSM では 12 個の初期値を指定している。AIC の算出においても未知パラメータの次元は 12 となるが，一方で変量効果 $\xi_{t1}, \xi_{t2}, \xi_{t3}$ は明かに定常性をもつため，散漫初期化された状態成分数は 6 のままとなる。

算出された AIC を前のモデルと比べると，この変量効果ありのモデルの AIC は

1753.4，前の変量効果なしのモデルの AIC は 1757.3 となり，変量効果ありの方がモデルの当てはまりが良いことが示された。

図 8 には，図 7 と同様に解析結果に基づく各年の 10 万人あたり死傷者数の期待値 $e^{\theta_{ti}}$ の推定値と，将来の予測値および予測区間を，それぞれ赤線と緑線で示している。予測が長期になるほど予測値と観測値の乖離が大きくなるのは図 7 と同じであるが，95%予測区間の幅が図 7 に比べてかなり縮まっていることがわかる。

図 7 の結果では観測モデルの分散が小さいために，観測値の増減のほとんどがトレンドの変化と見なされ，それゆえ長期予測においてトレンドのばらつきの大きさが予測区間長に影響したものと考えられる。それに対して，図 8 の結果では導入された変量効果が観測値の一時的な増減を吸収したため，赤線の推移が示すようにトレンドの変化が緩やかになり，長期予測においてもトレンドが保たれ予測区間長を狭めたものと考えられる。ポアソン分布の分散を超えるばらつき（過分散）を適切にモデルに導入することで，モデルの推定精度と予測精度はこのように逆に向上することとなる。

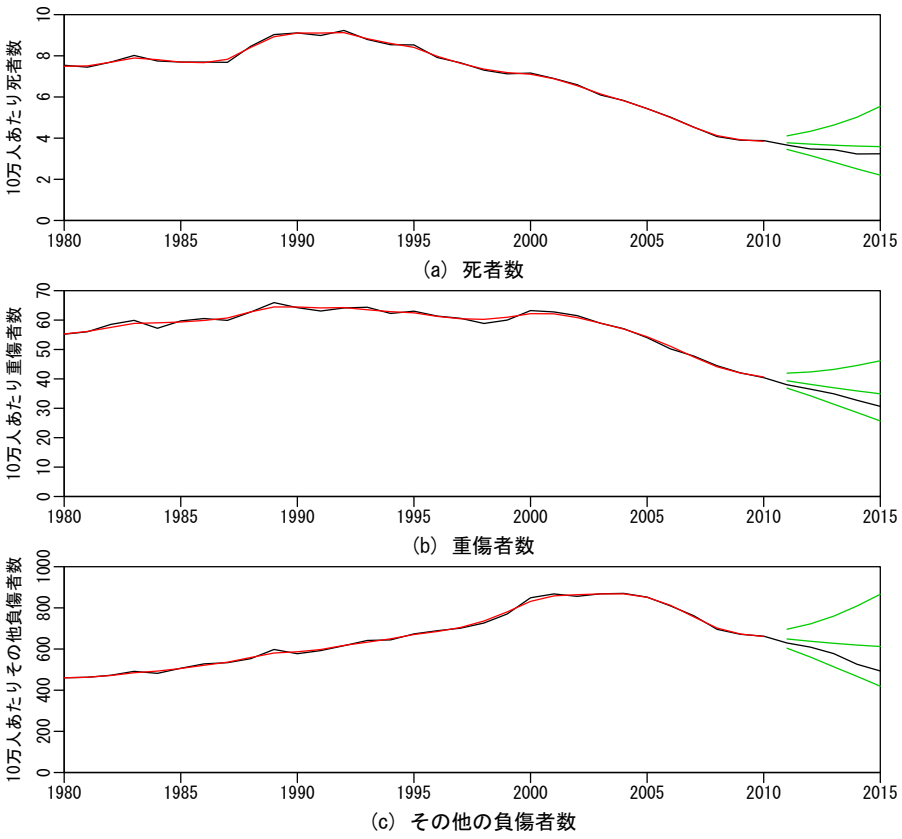

図 8　直近 5 年間の年間死傷者数の予測（変量効果ありのモデル）

6　結論

本稿では，事故件数のようなカウントデータの時系列を解析し予測するための線形非ガウス状態空間モデルを紹介し，統計解析ソフトRのパッケージ“KFAS”を用いた解析例を示した。はじめは月次推移データを扱い，トレンド成分と年周期の季節成分により将来の死者数について予測区間をもって適切に予測できることを示した。続いて，年間の死傷者数データについて，死者数，重傷者数，その他負傷者数の3変量による多変量時系列として，変数間の相関を組み入れたモデリング法を紹介し，さらに過分散を扱うための変量効果の導入により，予測区間が狭まってより精度よく予測できることを示した。

このように状態空間モデルを用いることで，将来の事故頻度だけでなく保険金単価も予測できるようになり，それらを織り込んだ料率算定や業績予想，リスク評価などが可能となる。将来予測を織り込んで算出された保険料率の例として，自然災害の長期的な増加トレンドを加味した火災保険の参考料率における長期係数が挙げられるが，状態空間モデルを導入すれば自然災害のトレンド変化を確率分布として見積もることができ，さらに毎年のリザルトを更新することでトレンドを随時修正することもできる。

本稿で示した解析例はリスクの区分数が3つと少ないが，複数のファクターをもつようなクラス料率算定では，多数ある料率クラスごとに変数をもつ大規模な多変量時系列を解析するため，状態空間モデルによるモデリングもより複雑になってくる。解析例で示したようにRパッケージ“KFAS”は多様な状態空間モデルを自由に組み合わせられる柔軟性をもっているが，将来的に現場のアクチュアリー実務に役立てられるためには，様々な状況に対応できる汎用的な状態空間モデルの設計手法の開発が必要になると考えられる。

謝辞

本稿の執筆にあたり，論文委員会より改訂のための有益なご助言を数多くいただいた。ここに，深く感謝申し上げたい。

参考文献

[1] Durbin J. and Koopman S. J. [2000],“Time Series Analysis of Non-Gaussian Observations Based on State Space Models from Both Classical and Bayesian Perspectives,” *Journal of Royal Statistical Society B*, **62**, pp.3-56.

[2] Durbin J. and Koopman S. J. [2012], *Time Series Analysis by State Space Methods* (2nd. Ed.), Oxford University Press.

[3] Helske J. [2016], “KFAS: Exponential family state space models in R,” Accepted to *Journal of Statistical Software*.

[4] 野村俊一 [2016], カルマンフィルター—R を使った時系列予測と状態空間モデル—, 共立出版.

Application of State Space Models to General Insurance : Exsamples of Analysis Using R Package “KFAS”

Shunichi NOMURA

Abstract

Insurance pricing methods have been developed over decades. To model claim count data, generalized linear models can treat discrete distributions such as Poisson distribution and estimate claim frequencies with respect to risk categories. On the other hand, because risks change as time passes, insurance pricing from recent claims experience do not always estimate future risks properly. It is important for management strategy, such as business planning and insurance pricing, to forecast time-varing risks. However, conventional linear regression model may be insufficient to forecast trend changes and so time series modeling is needed for proper prediction.

This paper introduces time series modeling for insurance claim frequencies via state space models. Linear state space models using non-gaussian distributions can be implemented easily by using a free statistical analysis software R with its package “KFAS”. Two illustrative examples of analysis for traffic accident casualties are shown with their source codes. Overdispersion for count data modeling is also discussed and incorporated into state space models in two ways.