統計数理(2011)
第 59 巻 第 2 号 181–192

特集「がん統計データおよびその解析」
[原著論文]

Transition model を用いた癌の死亡率データの解析及び予測について

緑川 修一[1] ・宮岡 悦良[2]

(受付 2011 年 1 月 4 日；改訂 9 月 15 日；採択 9 月 16 日)

要　　旨

がんの死亡者数やがんの罹患者数は 5 年もしくは 10 年毎に年齢階級別に発表されることが多く，このようなデータに対しての解析がよく行われてきた．本研究では，このようなデータに対して，解析及び短期予測を行うことを目的とする．本稿では厚生労働省データベースから，1950 年から 2000 年まで 5 年毎，15 歳から 90 歳以上の 5 歳階級からなるデータを用いた．このようなデータに対して，age-period-cohort model がよく用いられるが，パラメータの認定不可能の問題がある．Transition model を用いることで認定不可能の問題が解決でき，データへの当てはまりも age-period-cohort model に比べよいことが分かった．本研究では，時系列解析でよく用いられる state space model をデータに当てはめ，短期予測を行い，age-period-cohort model を用いた古典的な予測法との比較を行う．本研究では状態変数を求める際に particle filter を用いて解析を行い，予測分布を求め，予測分布の平均値を用いて，短期予測を行う．

キーワード： Transition model，age-period-cohort model，particle filter.

1. 導入

癌の死亡者数や罹患者数のデータは年齢階級，観測年別にまとまって発表されることが多くこのようなデータに対する解析及び予測がよく行われてきた．例えば，厚生労働省から発表されている癌の死亡者数のデータには表 1 のように，5 歳階級別，5 年毎に発表されているものがある．

ここで，観測年 t，年齢層 j における死亡者数を y_{jt}，人口を n_{jt} として表すものとする．観測年及び年齢層はそれぞれ $t=\{1,2,\ldots,T\}$，$j=\{1,2,\ldots,J\}$ グループ存在し，T と J は観測年及び年齢層の最大数を表すものとする．また，誕生群 k は $k=J-j+t$ と表せ，$k=\{1,2,\ldots,K\}$ グループ存在し，その最大数は $K=J+T-1$ となる．誕生群は例えば表 1 では，右下がりの対角線上に同一のグループが並ぶことになり同一グループに属する集団は誕生した年代が同じであると考えることが出来る．また，同一誕生群に属する集団は，例えば 1980 年に 45–49 歳の年齢層に属する集団は 1985 年には 50–54 歳の年齢層に属するため，各観測値が独立ではない．表 1 では，年齢層が 45–49 歳，50–54 歳，55–59 歳，60–64 歳の 4 グループ，観測年が 1980 年，1985 年，1990 年，1995 年，2000 年の 5 グループ，誕生群が $8=4+5-1$ の 8 グループからなる死亡者数及び人口を例として示した．

[1] セルジーン株式会社：〒100–0006 東京都千代田区有楽町 1–1–3
[2] 東京理科大学 理学研究科：〒162–0825 東京都新宿区神楽坂 1–3

表 1. 年齢層 j，観測年 t，に対する死亡者数 y_{jt} 及び人口 n_{jt}．$J=4$，$T=5$，$K=8$.

年齢層		観測年				
		1980 (t =1)	1985 (t =2)	1990 (t =3)	1995(t =4)	2000(t =5)
45 ∼ 49	誕生群	(k =4)	(k =5)	(k =6)	(k =7)	(k =8)
(j =1)	死亡者数	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
	人口	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}
50 ∼ 54	誕生群	(k =3)	(k =4)	(k =5)	(k =6)	(k =7)
(j =2)	死亡者数	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}
	人口	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}
55 ∼ 59	誕生群	(k =2)	(k =3)	(k = 4)	(k =5)	(k =6)
(j =3)	死亡者数	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}
	人口	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}
60 ∼ 64	誕生群	(k =1)	(k =2)	(k =3)	(k =4)	(k =5)
(j =4)	死亡者数	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}
	人口	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}

このようなデータに対しての解析は古くから行われており，Osmond (1985) はイングランドのウェールズ地方の肺癌の死亡率の解析を age-period-cohort model を用いて行っている．Age-period-cohort model は観測値 Y_{jt} が期待値 λ_{jt} のポアソン分布に従っていると仮定したモデルであり，以下のように表すことが出来る．

$$\Pr(Y_{jt}=y_{jt}|\lambda_{jt})=\begin{cases}\dfrac{\exp\{-\lambda_{jt}\}\lambda_{jt}^{y_{jt}}}{y_{jt}!} & y_{jt}=0,1,\ldots\\ 0 & \text{その他の場合}\end{cases} \tag{1.1}$$

$$\log(\lambda_{jt})=\log(n_{jt})+a+\alpha_j+\beta_t+\gamma_k\,.$$

ここで，a はモデルの切片，α_j，β_t 及び γ_k はそれぞれ，年齢層に対するパラメータ，観測年に対するパラメータ，誕生群に対するパラメータを表すものとする．Age-period-cohort model はデータに対しての当てはまりが良いことが知られており，良く用いられてきた．しかしながら，age-period-cohort model はパラメータの添え字間に $k=J-j+t$ という関係性が存在するために，パラメータを一意に推定出来ないという認定不可能の問題がある．Age-period-cohort model 及びその問題点について，詳細は Osmond and Gardner (1982)，Clayton and Schifflers (1987a, 1987b)，Holford (1991) に述べられている．

また，age-period-cohort model について，パラメータに制限を設けて認定不可能の問題を解決する方法や (Osmond and Gardner, 1982; Holford, 1991)，個々のパラメータは一意に推定出来ないが，推定可能なパラメータの関数の提案 (Clayton and Schifflers, 1987a; Holford, 1991) 等，様々な研究がなされている．丹後 (1985) では，パラメータの非線形成分が一意に推定出来ることを述べている．また，Midorikawa (2010) では mixed model を用いて認定不可能の問題を解決する手法が述べられている．本稿では，transition model を用いることで認定不可能の問題が回避出来ることを述べる．

ベイズの手法を用いた解析も行われており，Berzuini and Clayton (1994) や Knorr-Held and Rainer (2001) では各パラメータに事前分布を仮定した age-period-cohort を用いて肺癌の死亡率予測を行っている．また，Besag et al. (1995) は，ロジスティック回帰モデルを用いて癌の死亡率の解析を行っている．Bray et al. (2000)，Bray (2002) は Gaussian autoregressive prior model を用いてホジキン病の解析及び予測を行っている．本稿では，transition model を元にした予測法として，古典的な回帰分析を用いた予測法を行った．また，transition model の観

測年に対するパラメータに時系列構造を仮定したモデルを適用し，乱数を発生させる予測法を提案し，回帰分析を用いた予測法との比較を行う．

以下本稿では，同一誕生群に属する観測値がひとつ前の観測値に影響を受けると仮定した transition model について 2 章で述べる．Transition model のパラメータの推定には部分尤度関数を用いた最尤法を用いた．3 章では，日本の癌の死亡率データに対して transition model 及び age-period-cohort model を用いた解析結果について述べる．Transition model を元にした予測法について 4 章で述べ，5 章でまとめと今後の課題について述べる．

2. モデル及び推定法

2.1 Transition model

導入で示したように，本研究で扱うデータは互いに独立ではない．そこで本研究では以下の transition model を用いて解析を行った．Transition model は同一誕生群内のひとつ前の観測値が次の観測値に影響を与えると仮定したモデルであり，次のように表すことが出来る．

$$\Pr(Y_{kt'}=y_{kt'}|\lambda_{kt'},y_{kt'-1})=\begin{cases}\dfrac{\exp\{-\lambda_{kt'}\}\lambda_{kt'}^{y_{kt'}}}{y_{kt'}!} & y_{kt'}=0,1,\ldots\\ 0 & \text{その他の場合}\end{cases} \tag{2.1}$$

$$\log(\lambda_{kt'})=\log(n_{kt'})+a+\alpha_{g(k,t')}+\beta_{h(k,t')}+\gamma_k f(y_{kt'-1}).$$

ここで，k は誕生群に対する添え字，t' は同一誕生群の中で，観測された年代が古い順に $t'=1,2,\ldots,T'(k)$ と変化する添え字を表すものとする．t' の最大数は k により変化するため，$T'(k)$ と表すこととする．また，年齢層及び観測年に対する添え字は，k,t' が決まることにより一意に決まるので，関数 g と h を用いて $g(k,t')$, $h(k,t')$ と表すことができる．同一誕生群のひとつ前の観測値がどのように影響を与えるかの仮定を関数 f で定義することとする．

本稿では同一誕生群のひとつ前の観測値が次の観測値に影響を与えると仮定した 1 次の transition model を用いた．言い換えれば，同一誕生群のひとつ前の観測値を説明変数，次の観測値を被説明変数として用いたことになる．2 次やそれ以上の次数の transition model を用いることも可能であるが，transition model の次数を増やすと，被説明変数として用いることの出来る観測値の数が減ってしまうことになる．そこで本稿では 1 次の transition model を用いた．

Age-period-cohort model を用いた解析では，パラメータの添え字間に存在する線形関係とモデルの design matrix の要素が全て 0，1，もしくは -1 になることにより，認定不可能の問題が生じることになる．ここで， design matrix の要素に -1 が含まれるのは，年齢層，観測年，誕生群それぞれに 0 和制約を仮定した場合であり，カテゴリデータを説明変数として用いる際に良く用いられ，本論文で用いたモデルにおいても 0 和制約を仮定した．Transition model では，誕生群に対するパラメータに対応する design matrix の要素は 0，1 や -1 にはならない要素が存在する．これは，transition model は，観測値が同一誕生群の一つ前の観測値に影響を受けると仮定したためであり，同一誕生群の観測値が説明変数として，誕生群に対するパラメータに対応する design matrix に 1 や -1 の代わりに用いられるためである．そのため，認定不可能の問題を回避することが出来ると考える．$J=3$，$T=3$，$K=5$ の場合における $\boldsymbol{y}=(y_{11},y_{21},y_{31},\ldots,y_{13},y_{23},y_{33})^T$ に対応する design matrix の一部を，ひとつの例として以下に示す．

Age-period-cohort model

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}$$

Transition model

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & y_{00} & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & y_{10} & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & y_{20} & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & y_{01} & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & y_{11} & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & y_{21} & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & y_{02} \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & y_{12} & 0 \\ 1 & -1 & -1 & -1 & -1 & 0 & 0 & y_{22} & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{bmatrix}$$

ここで，design matrix 中の y も観測値と同様に，j と t を用いて y_{jt} で表すものとする．

2.2 部分尤度

Transition model のパラメータを推定する方法として，最尤法を用いた．尤度関数は次のように与えられる．

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \Pr(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{k=1}^{K} \Pr(\boldsymbol{y}_k|\boldsymbol{\lambda}_k), \tag{2.2}$$

ここで $\boldsymbol{\theta} = (a, \alpha_1, \ldots, \alpha_J, \beta_1, \ldots, \beta_T, \gamma_1, \ldots, \gamma_K)$, $\boldsymbol{y}_k = (y_{k1}, \ldots, y_{kT'(k)})$, $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kT'(k)})$ とする．Transition model は同一誕生群のひとつ前の観測値が次の観測値に影響を与えると仮定したモデルであるため，同一誕生群内において各観測値が独立ではない．各誕生群の観測値 $\boldsymbol{y}_k$ に対して，その結合確率分布は条件付き確率分布の定義より，

$$\Pr(\boldsymbol{y}_k) = \Pr(y_{kT'(k)}|y_{k1}, \ldots, y_{kT'(k)-1}) \Pr(y_{k1}, \ldots, y_{kT'(k)-1}),$$

となる．ここでモデルの仮定より，$y_{kT'(k)}$ は $y_{kT'(k)-1}$ にのみ依存するので，

$$\Pr(\boldsymbol{y}_k) = \Pr(y_{kT'(k)}|y_{kT'(k)-1}) \Pr(y_{k1}, \ldots, y_{kT'(k)-1}),$$

と表すことができ，$\Pr(y_{k1}, \ldots, y_{kT'(k)-1})$ に対して同様に条件付き確率分布及びモデルの仮定を用いて式変形し，$t' = 1$ まで続けると，

$$\left[\prod_{t'=1}^{T'(k)} \Pr(y_{kt'}|y_{kt'-1}, \lambda_{kt'}) \right] \Pr(y_{k0}),$$

と表すことが出来る．ここで，y_{k0} は誕生群 $k, t'=0$ における観測値を表すものとする．これを用いると尤度関数は，

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \Pr(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{k=1}^{K}\left[\prod_{t'=1}^{T'(k)} \Pr(y_{kt'}|y_{kt'-1},\, \lambda_{kt'})\right] \Pr(y_{k0}),$$

となる．ここで，尤度関数から $\Pr(y_{k0})$ を除いた，

$$PL(\boldsymbol{\theta}|\boldsymbol{y}) = \Pr(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{k=1}^{K}\prod_{t'=1}^{T'(k)} \Pr(y_{kt'}|y_{kt'-1},\, \lambda_{kt'}), \tag{2.3}$$

を部分尤度関数とし，部分尤度関数を最大にするパラメータの値を推定値とする．パラメータの推定に，y_{k0} は説明変数としてのみ用いるため，前述の観測値 $\boldsymbol{y}$ には含めていない．部分尤度関数から得られるパラメータの推定量の性質については，Kodem and Fokianos (2002) に述べられている．

3. データ解析

本章では，transition model，age-period-cohort model を用いたデータ解析について述べる．データは厚生労働省データベース (現在は政府統計の総合窓口 (e-Stat)) より，観測年が 1950 年から 2000 年までの 11 グループ，年齢層が 15–19 歳から 90 歳以上の 16 グループからなる，男性の胃癌の死亡者数のデータを用いた (http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do)．実データは，表 2 に掲載した．

Transition model は同一誕生群のひとつ前の観測値を説明変数として用いるため，観測値としては，1955 年から 2000 年までの 10 グループ，年齢層は 20–24 歳から 90 歳以上の 15 グループのデータを用いた．1950 年及び 15–19 歳の年齢層のデータは説明変数としてのみ用いた．

表 3 は，上記データに対して以下のモデルを当てはめた結果を示した．モデルは age-period-cohort model 及び transition model における関数 f を以下のように定義したモデルを用いた．

表 2. 胃癌による死亡者数 (人) - 男性 (http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do)．

年齢層	観測年										
	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000
15-19	9	9	15	17	19	9	5	5	11	5	4
20-24	27	30	46	72	80	59	28	21	18	20	18
25-29	65	106	127	158	162	162	104	87	49	49	30
30-34	166	196	300	353	346	309	308	196	102	77	70
35-39	359	395	470	615	628	562	526	453	315	207	142
40-44	788	854	790	781	1004	1003	799	719	646	494	303
45-49	1406	1568	1517	1309	1387	1638	1583	1192	1101	1027	724
50-54	2206	2470	2488	2402	1991	1922	2465	2203	1772	1626	1608
55-59	3024	3398	3717	3666	3214	2871	2632	3253	2992	2592	2458
60-64	3602	4125	4569	4993	4638	4201	3603	3467	4263	4034	3408
65-69	3465	4195	4799	5483	5699	5334	5013	4082	4081	5210	5237
70-74	2505	3244	4147	4483	5228	5594	5472	4952	4258	4869	6009
75-79	1063	1743	2289	3010	3459	4132	4743	4702	4613	4571	4859
80-84	261	479	829	1034	1457	2000	2636	3273	3512	4073	3977
85-89	61	78	170	241	313	536	804	1283	1727	2361	2767
90-	3	15	16	28	35	77	134	269	460	806	1184

表 3. それぞれのモデルに対する scaled residual SS，AIC 及びパラメータ数 p.

Model	Scaled residual SS	AIC	p
A-P-C	1253.525	2689.335	47
Transition	809.975	2226.887	48
Transition(rate)	404.267	1793.760	48

- Transition model (1): $f(y_{kt'-1}) = y_{kt'-1}$ とし，説明変数として観測値そのものを用いたモデル.
- Transition model (2): $f(y_{kt'-1}) = \frac{y_{kt'-1}}{n_{kt'-1}} 10000$ とし，人口 1 万人当たりの死亡者数を説明変数として用いたモデル.

関数 f は上記 2 つの関数以外にも様々な関数を適用することが可能である．本稿では，死亡者数が直接影響を与えると仮定した，直観的に分かり易い transition model (1) と，本稿で用いたデータへの当てはまりが特に良かった transition model (2) の結果のみ示すこととする．各モデルのデータへの当てはまりの良さを評価する指標として，

$$\text{Scaled residual SS(sum of square)} = \sum_{k,t'} \frac{(y_{kt'} - \hat{y}_{kt'})^2}{\hat{y}_{kt'}}$$

を用いた．また，age-period-cohort model，transition model (1) 及び transition model (2) を評価する指標として，

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2p,$$

を用いた．ここで，$\hat{y}_{kt'}$ は観測値 $y_{kt'}$ に対するモデルの当てはめ値，$\hat{\boldsymbol{\theta}}$ はパラメータの推定値ベクトル，$l()$ は対数尤度関数，p はパラメータ数をそれぞれ表すものとする．

Scaled residual SS は値が小さい程モデルのデータへの当てはまりが良いことが示され，AIC は値が小さい程良いモデル (真のモデルに近い) であることが分かる指標となる．表 3 より，データへの当てはまりが良いことが知られている age-period-cohort model よりも，transition model の方がデータへの当てはまりが良いことが見て取れる．このデータに対しては，3 つのモデルの中で transition model (2) がデータへの当てはまりが最も良いモデルであることが分かる．パラメータ数 p は，age-period-cohort model では，$J = 15, T = 10, K = 24$ で，年齢層，観測年，誕生群それぞれに 0 和制約を用いていることにより $p = 1 + (J-1) + (T-1) + (K-1) = 47$ として得られる．Transition model では，年齢層と観測年に 0 和制約を用いているため $p = 1 + (J-1) + (T-1) + K = 48$ として得られる．

4. Transition model を用いた予測法

この章では，transition model を用いた予測法及びその結果について示す．提案する予測法の精度を計るために，1955 年から 1995 年までの実データに対して transition model を当てはめ，2000 年の死亡者数を予測し，実データとの比較を行った．予測法として，回帰分析を用いた古典的な予測法と，観測年に対するパラメータに時系列構造を仮定した state space model を用い，得られた状態変数の分布から乱数を発生させることで予測値を得る方法の 2 つについて示す.

4.1 回帰を用いた予測法

$\tilde{\boldsymbol{y}}_{T+1} = (\tilde{y}_{1T+1}, \ldots, \tilde{y}_{JT+1})$ とし，データが観測年 T まで得られている時，$T+1$ の観測値を求めることを「1 期先予測する」こととする．回帰を用いて 1 期先予測を行うために，まず得

られた観測年 T までのデータに対して transition model を当てはめ年齢層，観測年及び誕生群に対するパラメータの推定値 $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{\alpha}_1, \ldots, \hat{\alpha}_J, \hat{\beta}_1, \ldots, \hat{\beta}_T, \hat{\gamma}_1, \ldots, \hat{\gamma}_K)$ を求める．年齢層 j の 1 期先予測値を $\tilde{y}_{jT+1} = \tilde{\lambda}_{jT+1} = n_{jT+1} \exp\{\hat{a} + \hat{\alpha}_j + \tilde{\beta}_{T+1} + \hat{\gamma}_k f(y_{j-1T})\}$ として求める．ここで，$\tilde{\beta}_{T+1}$ は一期先の観測年に対するパラメータを表すものとする．年齢層 j，観測年 $T+1$ の観測値 y_{jT+1} と同一誕生群のひとつ前の観測値は，j と t を用いて y_{j-1T} と表すことが出来る．1 期先予測値を求めるためには，$\tilde{\beta}_{T+1}$ が必要になる．$\tilde{\beta}_{T+1}$ は，得られた観測年に対するパラメータ $\hat{\beta}_1, \ldots, \hat{\beta}_T$ に回帰分析を適用して求める．本稿では直線回帰，

$$\hat{\beta}_t = a_0 + a_1 t,$$

及び 2 次の回帰曲線，

$$\hat{\beta}_t = a_0 + a_1 t + a_2 t^2,$$

を用いて，最小二乗法により a_0, a_1, a_2 を求め，予測値を得た．

図 1 (a) は，2000 年の日本の男性の胃癌での死亡者数(実測値)，と直線回帰を用いて予測した予測値(予測値 1)及び 2 次の回帰曲線を用いて予測した予測値(予測値 2)を横軸に年齢層，縦軸に死亡者数を配して図に表したものである．

どちらの予測値も実測値とは離れているが，2 次の回帰曲線を用いた予測の方が実測値に近いことが見て取れる．図 1 (b) は，同様にして女性の胃癌での死亡者数のデータを用いて予測

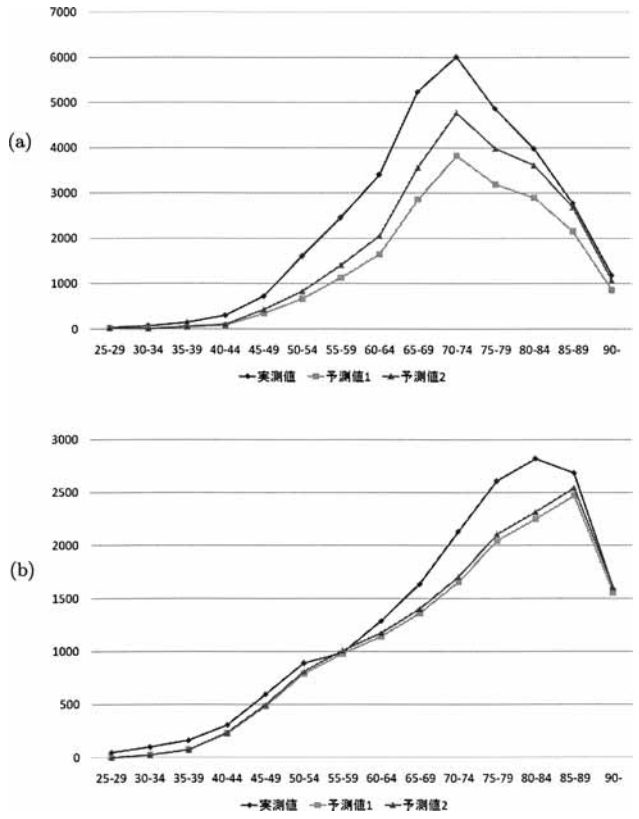

図 1. 回帰分析を用いて得られた予測値と実測値．胃癌男性 (a)，胃癌女性 (b)．

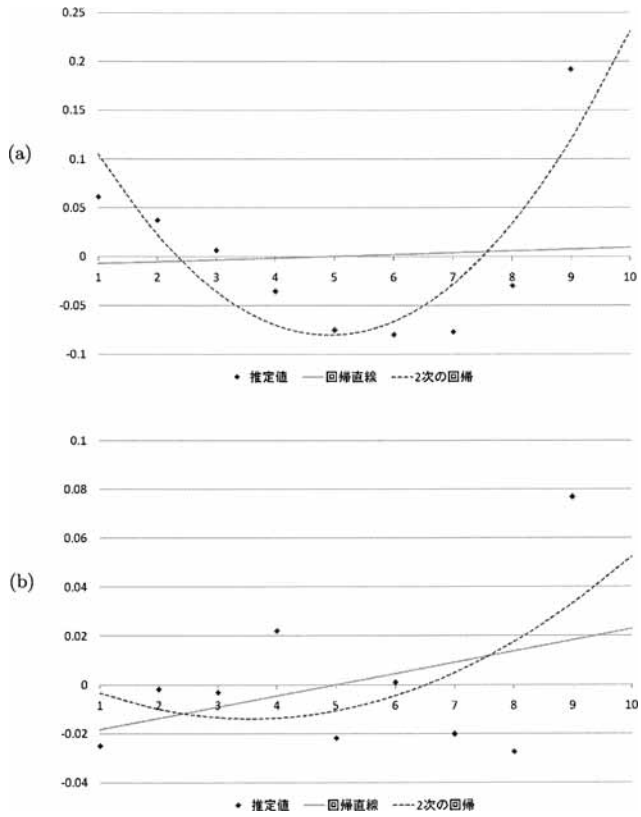

図 2. Transition model を用いて得られた観測年に対するパラメータの推定値と回帰直線及び 2 次の回帰曲線. 胃癌男性 (a), 胃癌女性 (b).

した予測値の図で, 男性のデータに比べ予測値 1 と予測値 2 に違いはあまり見られなかった. 参考として, transition model を用いて得られた観測年に対するパラメータの散布図, 回帰直線及び 2 次の回帰曲線を図 2 (a), (b) に示した. 図 2 (a) に比べ, (b) の方が $t=10$ において, 回帰直線と 2 次の回帰曲線の差が小さいことが分かる.

4.2 State space model を用いた予測法

次に, state space model を用いた予測法について述べる. 本研究で用いた state space model は, 観測年に対するパラメータがひとつ前の観測年に影響を受ける状態変数であると仮定したモデルで,

$$\Pr(Y_{jt}=y_{jt}|\lambda_{jt},\, y_{j-1t-1},\, x_t)=\begin{cases}\dfrac{\exp\{-\lambda_{jt}\}\lambda_{jt}^{y_{jt}}}{y_{jt}!} & y_{jt}=0,1,\ldots\\ 0 & \text{その他の場合}\end{cases}$$

$$\log(\lambda_{jt})=\log(n_{jt})+a+\alpha_j+x_t+\gamma_k f(y_{j-1t-1}). \tag{4.1}$$

$$x_t=x_{t-1}+\eta_t,$$

$$\eta_t\sim N[0,\tau^2].$$

と表すことが出来る. ここで x_t は観測年に対する状態変数で, η_t には互いに独立な, 平均 0,

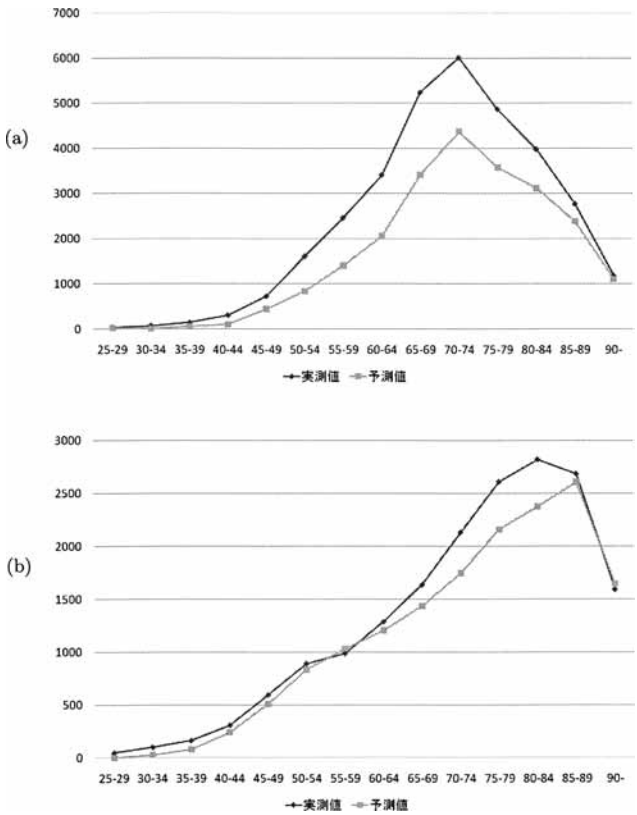

図 3. State space model を用いて得られた予測値と実測値．胃癌男性 (a)，胃癌女性 (b)．

分散 τ^2 の正規分布を仮定した．

State space model の解析方法として，particle filter (Doucet et al., 2001) を用いた．Particle filter のアルゴリズムを以下に示す．

Step1. $t=0$ とし，初期分布を設定し，$x_{0^*}^{(l)}, l \in \{1,\ldots,L\}$ を初期分布から L 個発生させる．
Step2. $t=t+1$ とし，$\eta_t \sim N[0,\tau^2]$ から乱数 $\eta_t^{(l)}$ を発生させ，$x_t^{(l)} = x_{(t-1)^*}^{(l)} + \eta_t^{(l)}$ を求める．
Step3. $p_l = \prod_{j=1}^{J} \Pr(y_{jt}|x_t^{(l)})$ を求める．
Step4. $(x_t^{(1)},\ldots,x_t^{(L)})$ から，確率 $\frac{p_l}{\sum_l p_l}$ で $(x_{t^*}^{(1)},\ldots,x_{t^*}^{(L)})$ を復元抽出する．
Step5. $t=T$ まで Step2 から Step5 を繰り返す．

Particle filter を用いて得られた $(x_{t^*}^{(1)},\ldots,x_{t^*}^{(L)})$ から，各 t の状態変数の分布を求め，他のパラメータは部分尤度関数を用いて求めた．また，初期分布の平均は age-period-cohort model を，1950 年から 2000 年までのデータに適用して得られた 1950 年の観測年に対するパラメータの推定値を用いた．分散については，どのような値を仮定して良いかの明確な指標がないため，いくつかの値を用いて解析を行った．分散が大きすぎると，各粒子に対して，復元抽出する際に求める確率がほとんど 0 となってしまう．そこで本論文における解析では，大きな値からはじめ，復元抽出の確率がほとんど 0 とならないように値を徐々に小さくしていき，最終的に分

表 4. それぞれの予測法に対する scaled residual SS (胃癌男性).

Model	予測法	Scaled residual SS
Transition	直線回帰	11274.223
Transition	2 次の回帰曲線	4831.437
State space	乱数を発生	5792.206

表 5. それぞれの予測法に対する scaled residual SS (胃癌女性).

Model	予測法	Scaled residual SS
Transition	直線回帰	1982.518
Transition	2 次の回帰曲線	1809.126
State space	乱数を発生	1687.901

散を 0.1 と仮定して解析を行った.

State space model を用いた予測法を以下に示す. 回帰分析を用いた予測法と同様に, 観測年 T までのデータに state space model を当てはめ, 得られた観測年に対する状態変数 $x_{T*}^{(l)}$ 及び $\eta_{T+1} \sim N[0, \tau^2]$ から乱数 $\tilde{\eta}^m$ を発生させ,

$$\tilde{x}_{T+1}^m = \frac{1}{L} \sum_l x_{T*}^{(l)} + \tilde{\eta}^m ,$$

を求める. 得られた $\tilde{x}_{T+1}^m$ を用いて x_{T+1} の予測分布を求める. 本研究では, $M = 1000$ とし, $\frac{1}{M} \sum_m \tilde{x}_{T+1}^m$ を用いて予測を行った. $\eta_{T+1} \sim N[0, \tau^2]$ としているため M の数を多くすると $\frac{1}{L} \sum_l x_{T*}^{(l)} \approx \frac{1}{M} \sum_m \tilde{x}_{T+1}^m$ となる事に注意されたい. また, 年齢層に対するパラメータ及び誕生群に対するパラメータは, 回帰分析を用いた予測法と同様に仮定し予測を行った.

図 3 (a) は, 2000 年の日本の男性の胃癌での死亡者数(実測値) 及び state space model を用いて得られた予測値(予測値) を横軸に年齢層, 縦軸に死亡者数を配した図である. 図 1 (a) の直線回帰を用いた予測値(予測値 1) よりも実測値に近く, 2 次の回帰曲線を用いた予測値(予測値 2) より実測値から離れていることが見て取れる. 図 3 (b) は, 同様にして女性の胃癌での死亡者数のデータを用いて予測した予測値の図で, 図 1 (b) に比べ若干ではあるが実測値に近い予測値が得られた. それぞれの予測法に対する scaled residual SS を, 表 4 及び表 5 に示した.

5. まとめと今後の課題

本研究で用いた, 年齢階級別にある程度まとまって発表されるようなデータに対して, age-period-cohort model はよく用いられてきた. Age-period-cohort model はデータへの当てはまりがよく, 年齢層と観測年の影響のみを仮定した age-period model, 観測年の影響として観測年 t の関数を用いたモデルや mixed model に比べデータへの当てはまりがよいことが Midorikawa (2010) で述べられている. 本論文で用いたデータは Midorikawa (2010) で用いられたデータと同じであり, transition model を用いた解析では age-period-cohort model を用いた解析よりもデータへの当てはまりがよいことが分かった. 特に, 説明変数にひとつ前の観測値を元にした死亡率を用いたモデルでは, 他のモデルに比べてデータによく当てはまっていることが分かった. また, transition model には age-period-cohort model にある認定不可能の問題がなく, パラメータを一意に推定することが出来る.

予測に関しては, データへの当てはまりのよい transition model を元にした 2 つの手法を用いて予測を行ったが, どちらの手法もあまりよい予測が行えなかった. 回帰分析を用いた予測

法では，どの回帰モデルを用いるかの選択があり，state space model を用いた予測法でも 1 次のトレンドモデル，2 次のトレンドモデル等の選択がある．どの手法を用いて予測を行えば良いかの明確な指標はなく，その選択は解析者に委ねられることになるため注意が必要である．State space model を解析する際に用いた particle filter は実装するのが容易であり，予測する際のひとつの選択肢として有用であると思われる．

本論文では予測する際に，予測値の新しい誕生群に対するパラメータを欠損として扱い予測値を算出しなかったが，予測する期間が増えれば増えるほど欠損値が増え，予測として機能しなくなってしまう．誕生群に対するパラメータの予測をどのように行っていくかが，今後の課題のひとつとして挙げられる．また，本論文ではあまり取り上げなかったが，人口の推移も予測値に多大な影響を与えるものと考えられる．人口の予測も同時に行うような手法を本論文で取り上げたデータに応用することで，予測に対する改善がみられるのではないかと考えている．

謝　辞

本論文改訂にあたり，匿名のお二人の査読者から非常に有益なご意見を頂いた．ここに感謝の意を表したい．

参 考 文 献

Berzuini, C. and Clayton, D. (1994). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine*, **13**, 823–838.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science*, **10**, 3–66.

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality, *Journal of the Royal Statistical Society. Series C*, **51** (2), 151–164.

Bray, I., Brennan, P. and Boffetta, P. (2000). Projections of alcohol- and tobacco-related cancer mortality in central Europe, *International Journal of Cancer*, **87**, 122–128.

Clayton, D. and Schifflers, E. (1987a). Models for temporal variation in cancer rates II: Age-period-cohort models, *Statistics in Medicine*, **6**, 469–481.

Clayton, D.and Schifflers, E. (1987b). Models for temporal variation in cancer rates I: Age-period-cohort models, *Statistics in Medicine*, **6**, 449–467.

Doucet, A., de Freitas, N., Gordon, N. and Editors. (2001). *Sequential Monte Carlo Methods in Practice*, Springer, New York.

Holford, T. R. (1991). Understanding the effects of age, period and cohort on incidence and mortality rates, *Annual Review of Public Health*, **12**, 425–457.

Knorr-Held, L. and Rainer, E. (2001). Projection of lung cancer mortality in West Germany: A case study in Bayesian prediction, *Biostatistics*, **2**, 109–129.

Kodem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, New York.

Midorikawa, S. (2010). Application of mixed model to the cancer mortality data in Japan, *JP Journal of Biostatistics*, **4**, 13–31.

Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates, *International Journal Epidemiology*, **14**, 124–129.

Osmond, C. and Gardner, M. J. (1982). Age, period and cohort models applied to cancer mortality rates, *Statistics in Medicine*, **1**, 245–259.

丹後俊郎 (1985). 年齢・時代・コホートの効果の推定―線形成分と非線形成分への分解―，応用統計学，**14**, 45–49.

Application of Transition Models to Cancer Mortality Data

Shuichi Midorikawa[1] and Etsuo Miyaoka[2]

[1]Celgene K.K.
[2]Department of Mathematics, Tokyo University of Science

Cancer mortality data are often published as specific statistics such as a table of age- and period- by public institutions. In this study, we focus on analysis and prediction of cancer mortality using data collected by MHLW (Ministry of Health, Labour and Welfare). Generally an age-period-cohort model is used in the analysis of these data, but this model has a problem of being non-identifiable due to over parameterization. Applying a transition model to these data overcomes that problem and makes it fit well. In this study we predict cancer mortality applying a state space model which is used in a time series analysis, and we compare our prediction method with the classical prediction method of the age-period-cohort model. We use the particle filter for the state space model, and we predict the death count by prediction distribution.

Key words: Transition model, age-period-cohort model, particle filter.