厚生労働科学研究費補助金（がん対策推進総合研究事業）
（分担）研究報告書

罹患数の遅れ報告補正に関する統計学的手法の検討

加茂 憲一　　札幌医科大学　医療人育成ｾﾝﾀｰ　准教授

研究要旨 がん罹患数に関する情報は、毎年の公式報告（同年のMCIJ）以降も繰り返し収集されている。この繰り返し収集の情報と当初の公式報告との間に差異が無ければ問題ないが、実際には初期報告より多い数が報告される傾向にある。この事実を踏まえ、本研究では罹患の追加報告に関する時系列の挙動に着目し、追加報告される数に関する挙動のメカニズムを数理モデルによって表現することを試みた。このモデルに実データを適用することによって、最終的な追加報告の漸近値に対する見通しを示すことが可能となった。

A. 研究目的

都道府県におけるがん罹患数の情報は、MCIJ（Monitoring of Cancer Incidence in Japan）として毎年集計・報告されている。例えば2003年罹患は、MCIJ2003として、2008年3月に報告されている。一方で、その後の2004年罹患を集計するMCIJ2004（2008年12月収集）に際しては、2004年罹患のみならず2003年罹患も再報告される。その後も引き続き、2005年罹患、2006年罹患…が報告される際にも2003年罹患は再報告され続ける。その理由は、MCIJ2003として2003年罹患が報告された以降にも、2003年罹患に関する修正報告がなされる可能性があるため、アップデートする必要があるからである。このように、該当するMCIJ年以降に報告されるケースを本報告書では「遅れ報告」と呼ぶこととする。

遅れ報告の発生は、主に以下の3つが考えられる：

１．追加登録

２．登録削除

３．属性変更

１の追加登録とは、MCIJ報告の締切に間に合わなかったケースが、後年に追加されてくる案件である。この場合、該当年の罹患数は増加する。２の登録削除は、後に別年の罹患であることが判明した等の理由により、該当年のレコードから削除されるケースである。この場合、該当年の罹患数は減少する。３の属性変更は、罹患者の情報に誤りがあった場合、例えば登録年齢に誤りがあった場合に、そのケースの属性を変更するものである。この場合、該当年の罹患数は変わらないが、該当年内での情報が変更される。日本における遅れ報告の大部分は１の追加情報が占める傾向が強いため、遅れ報告における変更は追加登録に集約されるものと仮定して議論を進める。

公式な罹患報告以降に数値がアップデートされる問題に対して、統計学的アプローチを適用した先行研究としては、隣り合う

年の罹患数の比に対するMANOVA（多変量分散分析）モデルによるアプローチがNCIのホームページで紹介されている（https://surveillance.cancer.gov/delay/methods.html）。この手法を日本におけるがん登録データに適用した結果については２年前の報告書に記載した。MANOVAモデルにおいて問題となるのは、罹患数の「比」を被説明変数と設定しているため、追加報告が時間∞において有限な漸近値を持つことなく増え続ける点である。つまり、本研究テーマの目的である「将来分を積み上げた最終的な罹患数を明らかにする」ことが、MANOVAモデルでは達成されない。一方で別のアプローチとして、遅れ報告の増分と減分を分割し、二項分布とポアソン分布の混合によってモデルを構築するアプローチも存在する（Douglas et al. J Am Stat Assoc, 2005）。統計学的な手法としては現時点においてこの方法がベストと考えられるが、この手法では遅れ報告が過去情報を含む履歴として保存されている必要がある。通常の登録修正は上書きされることが多いため、このモデルを日本データに適用することは一部のがん登録データに限られる。そこで本報告では、遅れ報告の積み上げが時間∞で有限な漸近値を有することを前提とした数理モデルを再構築し、その手法を愛知県のがん登録データに適用した結果を報告し、今後の発展可能性について議論する。

B. 研究方法

遅れ報告の特性をモデル化するにあたって、現行のMANOVAモデルが抱える問題点である、遅れ報告の積み上げが時間∞において有限な漸近値を持たない点に着目し、この問題点を改良する新たな数理モデルを構築する。遅れ報告の時間と共に増加傾向にあるという特徴から、数値は積み上がることとなり、その積み上げは漸近値を有する必要がある。そのような状況に数学的な表現を適用すると、時間変化に対応する数列における無限級数が収束するという現象に酷似している。無限級数が有限な漸近値を有する数列として代表的なのは等比数列であるため、遅れ報告のメカニズムが等比数列に従うものと仮定して数理モデルを構築する。

$I_{m,n}$ を、m年罹患のMCIJnにおける報告数とおく。前述の通り、$I_{m,\infty}$は有限な値に収束することが期待される。そのためにはmを固定する毎に$\{I_{m,n} - I_{m,n+1}\}$が$n\to\infty$において0に収束する事が必要条件である。このことを踏まえて、x_kを$I_{m,m+k+1}$と$I_{m,m+k}$との差分、すなわち

$$x_k = I_{m,m+k+1} - I_{m,m+k}$$

とおき、数列$\{x_k\}$が等比数列に従うと仮定する。すなわち、差分x_kを

$$x_k = x_1 \times r^{k-1}$$

と設定する。ここで、無限級数 $\Sigma^{\infty} x_k$ は収束する必要があるため、公比rには0より大かつ1未満との制約を与える。この制約がみたされる場合、$\{x_k\}$の無限等比級数は

$$\Sigma^{\infty} x_k = x_1 / (1-r)$$

と有限な値に収束するため、m年罹患数に対する報告遅れ∞での漸近値$I_{m,\infty}$は

$$I_{m,\infty} = I_{m,m} + x_1 / (1-r)$$

で推定される。ここで必要なのは、公比パラメータrを実際のデータから推定することであり、最小二乗推定を適用する。

公比パラメータ推定においては、MCIJとして収集されたデータを用いる。利用する

データの状況を表１に示す。罹患年としては1993年～2015年罹患（表１の行に対応）が、MCIJ2003～2015（表１の列に対応）の際に収集されている。表中の数値は、リアルタイムなMCIJからの遅れ年を表しており、0の箇所の罹患数がMCIJとして公式報告されている。尚、空欄の箇所にはデータが存在しない。

表１　MCIJ収集データの状況

		MCIJ												
		2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
罹患年	1993	10	11	12	13	14	15	16	17	18	19	20	21	22
	1994	9	10	11	12	13	14	15	16	17	18	19	20	21
	1995	8	9	10	11	12	13	14	15	16	17	18	19	20
	1996	7	8	9	10	11	12	13	14	15	16	17	18	19
	1997	6	7	8	9	10	11	12	13	14	15	16	17	18
	1998	5	6	7	8	9	10	11	12	13	14	15	16	17
	1999	4	5	6	7	8	9	10	11	12	13	14	15	16
	2000	3	4	5	6	7	8	9	10	11	12	13	14	15
	2001	2	3	4	5	6	7	8	9	10	11	12	13	14
	2002	1	2	3	4	5	6	7	8	9	10	11	12	13
	2003	0	1	2	3	4	5	6	7	8	9	10	11	12
	2004		0	1	2	3	4	5	6	7	8	9	10	11
	2005			0	1	2	3	4	5	6	7	8	9	10
	2006				0	1	2	3	4	5	6	7	8	9
	2007					0	1	2	3	4	5	6	7	8
	2008						0	1	2	3	4	5	6	7
	2009							0	1	2	3	4	5	6
	2010								0	1	2	3	4	5
	2011									0	1	2	3	4
	2012										0	1	2	3
	2013											0	1	2
	2014												0	1
	2015													0

表１に示されたデータは都道府県規模で集約されているが、全ての都道府県において表１が完全に充たされている訳ではない。この表が完全に充たされているのは、山形県、新潟県、福井県、愛知県、滋賀県、長崎県、熊本県の７県であった。解析結果の安定性を考慮し、この７県の中から最も人口規模の大きな愛知県を選択し、解析を行った。

実解析においてはソフトウェアR（ver.4.0.3）を用いた。パラメータ推定においてはoptimize()関数を用い、公比パラメータの範囲を(0,1)に制限した推定を行った。

C. 研究結果

愛知県の男性に関する罹患情報を用いた解析結果を紹介する。図１は、愛知県の男性に関する罹患数に対する経年変動および遅れ報告の挙動をヒートマップで表したものである。縦軸が罹患年、横軸がMCIJ報告からの遅れ年を表し、この組み合わせにおける罹患数が色の濃淡（濃色が高く薄色が低い）と等高線（線上の数値が罹患数）によって表現されている。左上の空白領域は情報が存在しない箇所であり、表１の左下領域に対応している。ヒートマップの縦方向は罹患の経年的な変動を、横方向は遅れ報告による変動を表している。ヒートマップの上側になるほど罹患数が高い傾向を示していることから、愛知県における罹患数は経年的に増加傾向であることが分かる。また、等高線が横ばい傾向であるのは、縦方向の変動が大きい、つまり時系列の罹患数の変動に比して遅れ報告の割合が低いことを意味している。しかし、横ばい傾向の等高線を注意深く観察すると、僅かながら右下がりの傾向も観察されることから、遅れ報告は基本的に増加傾向にあることも分かる。更に右下がり傾向の度合いについては、初期段階（空白領域との境界近く）で強い一方で、初期段階から離れるに伴って真横に近くなる傾向も観察される。これは、遅れ報告がいずれゼロに漸近することを示唆するものである。

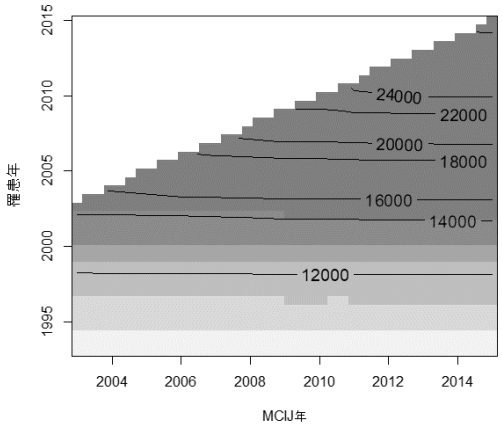

図１　愛知県（男性）の罹患数および遅れ報告

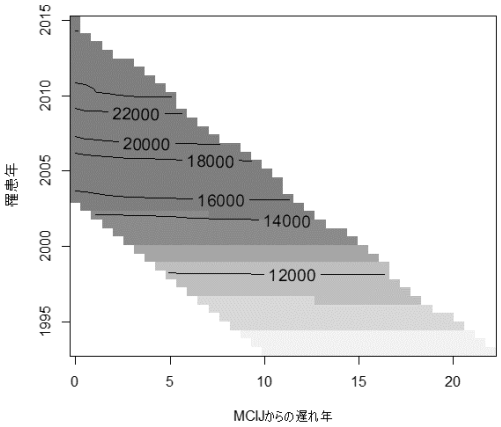

図２　報告遅れ年による罹患数の変動

図１における横軸はMCIJ年であるが、本研究では遅れ年による影響をターゲットとしていることを踏まえ、ヒートマップの横軸を図１のMCIJ年から「MCIJ年からの遅れ年」に設定し直したのが図２である。図１の空白領域が、図２における直線y=2003より上側の空白領域に対応する。また、右下の空白領域は図１と同じく、罹患年＋遅れ年が最新MCIJ年を超えるためデータが存在しない領域である。左下の空白領域は、今回解析に用いた表１のデータには含まれなかったが、実際には過去分の情報が存在する可能性がある領域である。図２の表現により、罹患年は異なるが遅れ年が同じデータが縦方向に揃い、MICJ年の直近が左側に揃った。このことにより、図１で確認された「初期段階での右下がり傾向が他より強い」という性質が左端に揃うことにより再確認できる。

図１と２において観察された「MCIJからの初期段階においてヒートマップの等高線の右下がり傾向が強い」点について、経年的なボックスプロットを用いて再確認したのが図３である。縦軸が隣り合う報告年における罹患数の差分（新しい報告年から古い報告年を引いたもの）を表し、横軸がMCIJからの遅れ年を表している。横軸の始点1は、n年罹患に対するMCIJn年報告と、その翌年のMCIJ(n+1)年において報告されるn年罹患数の差（例えば2003年罹患に対するMCIJ2004での報告数とMCIJ2003での報告数の差）を意味する。まず、数値の大半が正値であるので、遅れ報告は基本的に追加報告であることが分かる。次に、初期段階から平均的に減少傾向で0に近付くこと、それに伴って分散も減少傾向であることが観察される。(1,1099)に観察される高い外れ値は、2010年罹患に対するMCIJ2011報告とMCIJ2010報告の差分である。

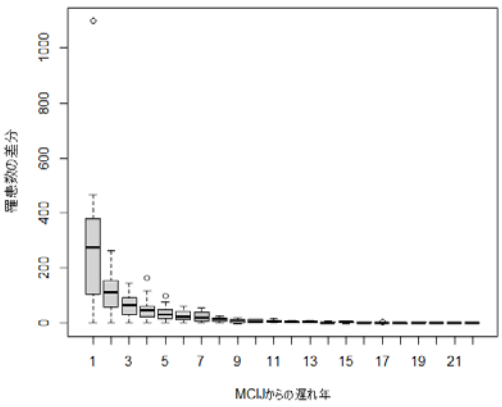

図３　隣り合う報告年における罹患数の差分

本研究では、隣り合う報告年における罹患数の差分を、等比数列を用いてモデル化する。そこで、罹患数の差分に関する傾向を、図２と同じ座標軸を用いたヒートマップにより表現したのが図４である。左側に高い数値が固まっていることから、主たる遅れ報告はMCIJ報告の直後に発生していることが分かる。その中でも、2010年罹患の初期段階周辺において、特異的に差分の大きな箇所が観察されるが、これは図３の(1,1099)に対応する。

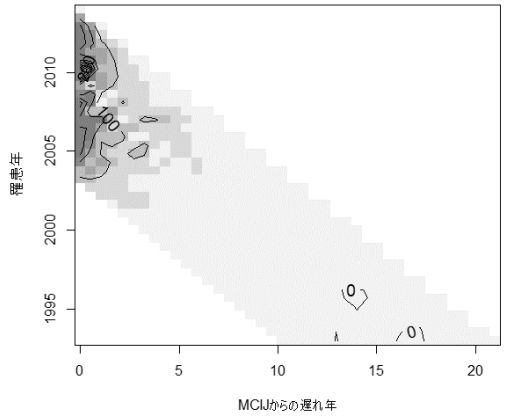

図４　差分に関するヒートマップ

図１～４に基づくデータの全体傾向を踏まえ、罹患年を固定した時系列解析を試みる。本研究において用いるモデルは等比数列（隣り合う罹患年の差分に等比数列を仮定する）に基づいている。等比数列においては、初項と公比が必須である。このうち、公比はパラメータとして実データから推定することになるが、初項が得られるのは表１において0と1が揃っている罹患年ｍｐ、おである。更に表１において2が存在しない罹患年は初項と第2項のみから公比を推定することとなり、公比の推定値がx_2/x_1となるため事実上無意味である。従って、2003年罹患～2013年罹患のみが解析に利用できるデータとなる。表１においてMCIJに関しては2015年まで存在することから、上記の罹患年において遅れ報告の情報量が最も多く蓄積される2003年罹患（13年分の情報を有す）に着目した解析を試みた。2003年罹患は、図１、２、４において、縦軸が2003の箇所から真横に引いた直線上の数値に対応する。

まず、2003年罹患が遅れ報告によりどのように変化するかを図５に示す。横軸はMCIJ報告からの遅れ年（始点の0はMCIJ2003に対応）を、縦軸は罹患数を表す。15000人規模の罹患数に比すると僅かではあるが、12年で400人弱の追加報告が観察された。経年的な挙動としては遅れ年に伴って単調増加であるが、その増加傾向（傾き）は遅れ年が進むに伴って緩やかになることが観察された。

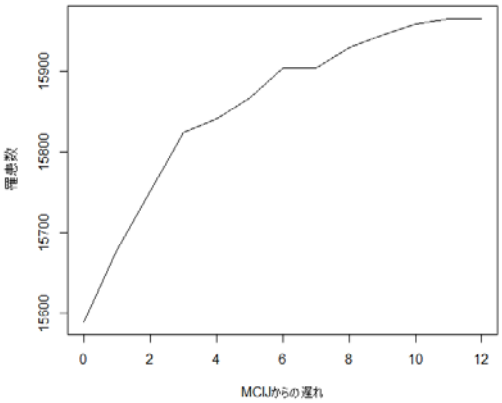

図５　2003年罹患（愛知県・男性）

本研究においては、隣り合う報告年の差分を等比数列によってモデル化する。図５の情報を、隣り合う遅れ年に対する差分により表現し直したのが図６であり、図４において直線y=2003上の値に対応している。縦軸が罹患数の差分、横軸がMCIJからの遅れ年を表している。横軸の始点1は、リアルタイムのMCIJ年（MCIJ2003）とその翌年（MCIJ2004における2003年罹患）の差に対応している。図６から、罹患数の差分に関する経年的な減少傾向が観察される。従って、このデータに対して等比数列を適用すると、0より大かつ1未満の公比が推定されることが期待される。このことは、差分がいずれ0に収束し、罹患数が漸近的に有限値に収束することを意味する。

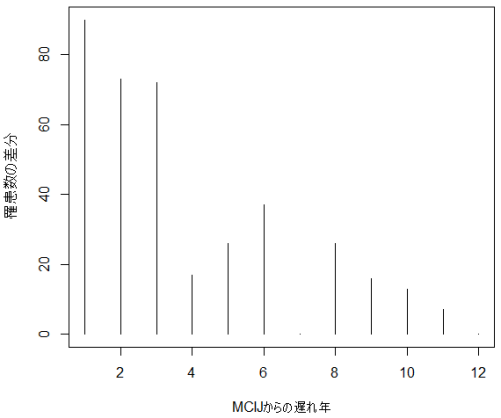

図６　2003年罹患の罹患報告年に関する差分

図６に対し、前述の等比数列モデルを適用し、最小二乗法を用いると、公比パラメータは0.76762と推定された。この公比パラメータと差分の初期値（図６の横軸＝1に対する差分）を用いると、差分の期待値が帰納的に算出される。その結果を図７に示す。これは、図６の実測値に対して、推定された差分の期待値を曲線で重ね描きしたものである。尚、初項は実測値をそのまま用いるので、実測値と期待値は必ず一致する。

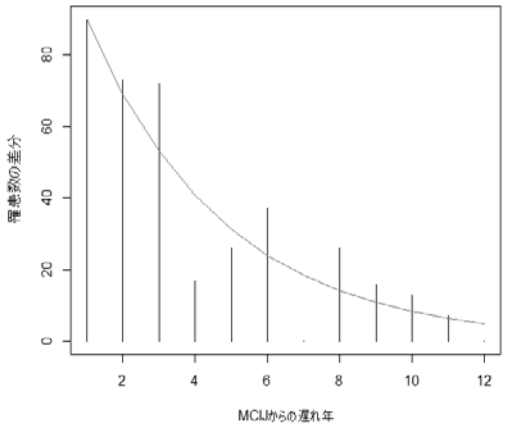

図７　罹患数の差分の推定結果

罹患数の差分に対する公比パラメータの推定結果（0.76762）から算出される差分の期待値、および初年の罹患数の実測値を用いれば、その後の罹患数の期待値が帰納的に再現できる。その結果を図８に示す。オーバープロット（図５の折れ線にプロットを加えたもの）が罹患数の実測値、プロットの無い折れ線が期待値を表す。差分に対する等比数列の初項に実測値を適用することから、罹患数に関して最初の２年分については実測値と期待値は一致し、３年目以降が事実上の推定値となる。

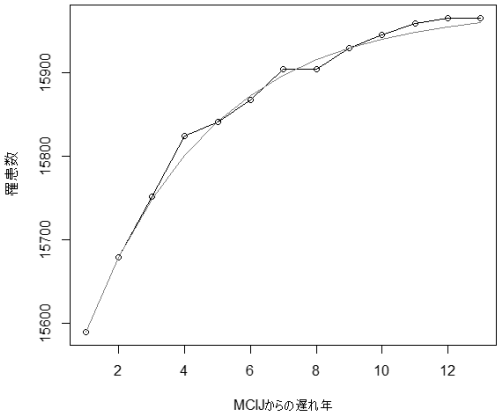

図８　罹患数の推定結果

D. 考察

MCIJ報告以降にアップデートされるがん罹患数に関する情報（遅れ報告）に着目し、それを数理モデルによって表現することを試みた。遅れ報告は基本的に追加報告であると仮定し、その増分が等比数列に従うモデルを構築した。モデルに含まれる唯一のパラメータは等比数列における公比であり、それを最小二乗法により推定した結果、2003年における愛知県の男性罹患数に関しては0.76762と推定された。図８は、推定された結果を実測値に重ね描いたものであるが、実測値への適合状態は悪くなく、差分の等比数列に基づくモデリングが十分に機能していることが伺える。

今回等比数列を適用したのは、罹患数を有限な漸近値に収束させることが最大の理由である。実際に公比は0.76762と0より大かつ1未満の値が推定されたため、差分の無限級数は有限な値に収束する。具体的には $I_{2003,\infty} = 15589 + 90/(1\text{-}0.76762) = 15976$人が愛知県における男性の2003年罹患数の報告遅れ時間∞に対する漸近値となる。上式中の15589はMICJ2003における2003年罹患、90は差分の初年値（等比数列の初項＝15679人－15589人）、0.76762は公比パラメータの推定値である。2003年罹患の漸近値が15976人と推定された結果は、MCIJ2003における15589人という報告に対して、その後387人の追加が見込まれることを示唆している。また、MCIJ2015の際に収集された2003年罹患数は15966人であることから、12年経過すればMCIJ2003報告以降に見込まれる増分の約97%が既に充足されていることが予測される。

E. 結論

本研究では、毎年報告されるがん罹患数について、MCIJ以降に追加報告される事案に着目した。追加報告を等比数列によってモデル化し、実データ（愛知県の2003年男性の罹患）を用いた検証を行った。追加報告に対するフィッティングは悪いものでなく、本モデルが機能していることが伺える。本モデルは、最も情報量の多い2003年罹患に

限定しての解析であったが、今後は複数年の罹患情報を用いて都道府県独自のパラメータを推定するという発展性が考えられる。他にも考えられる改良点や、現行モデルの有する問題点も存在するので、本研究テーマの今後の発展に資することを期待し、それらを以下に列挙しておく。

まず、データに依存する問題点として、罹患年とMCIJ年との間の乖離幅が一定でない点が挙げられる。今回解析に用いたデータ（MCIJ2003～2015）の収集されたタイミングを表２に示す。例えば、解析対象とした2003年罹患はMCIJ2003として2008年3月に収集されているため、4年3カ月のラグが存在していた。このようなラグは年々改良され、その後2015年罹患がMCIJ2015として収集されたのは2008年9月であったため、この時のラグは2年9カ月である。報告の即時性という観点からは、このラグは短い方が理想的であるが、ラグを縮めた皺寄せは遅れ報告の増加という形で発生する可能性が高い。12年でラグを1年6カ月縮めてきた影響が、遅れ報告の増加に転嫁されている可能性があるが、現行のモデルではその調整ができない点が問題である。また、このラグが縮まるプロセスも連続的に改良されてきた訳ではない。特に、2009年罹患と2010年罹患は同じタイミング（2013年9月）に収集されている。つまり、ある年の罹患数が、MCIJ2009とMCIJ2010で同数となり、このタイミングで差分は必ずゼロとなる（図２における横軸=7の部分）。ここで特異な差分が観察されること、そしてその影響が前後の年に波及する（図４の(0,2010)周辺に観察されるホットスポット）ことは十分に考えられるが、この点に関して現行のモデルでは調整ができない。特に2009年罹患に関しては差分の初項がゼロとなり、等比数列が無意味となってしまい、本手法を適用することができない。初項がゼロにならずとも、本モデルでは差分を積み上げる形で罹患数を推定するため、初項近辺に特異な項が位置する場合には本手法による推定が機能しない危険性もある。このように、MCIJ収集のタイミングや方法が一様でないことをモデルとして表現できない点、そして初項近辺に特異な挙動をする観測年に対して頑健でない点がデータに依存する問題点である。

表２　MCIJの収集タイミング

MCIJ年	報告年月	備考
2003	2008年3月	
2004	2008年12月	
2005	2009年9月	
2006	2010年9月	
2007	2011年9月	
2008	2012年9月	
2009	2013年9月	この2年は報告年月が同じ
2010	2013年9月	
2011	2014年9月	
2012	2015年9月	
2013	2016年12月	一部の地域が2017年3月
2014	2018年1月	一部の地域が2018年3月
2015	2018年9月	一部の地域が2018年10月

次に、モデル依存の問題点に着目する。本研究では連続する年の間における罹患数の差分が等比数列に従うことを前提としている。等比数列は初項と公比が必要であるが、差分をターゲットにしているため、差分の初項が必須となる。このことはMCIJ年に加え、その翌年の情報が無ければ今後の推定ができないことを意味する。つまりMCIJの

翌年になって初めて推定が可能になるという即時性に関する問題点が存在する。もう一つのモデル依存の問題点として考えられるのは、公比パラメータを一様に設定している点である。例えば、先行研究であるMANOVAモデルにおいては、罹患数の比が遅れ年によって４パターンに分類されるものと仮定し、４パターンのパラメータを推定している。このことにより、変化の割合が途中で変わる場合にも対応が可能となるが、本モデルは無限級数を最終的なアウトプットとしているため、途中で公比を変えることは現実的でない。現時点での観察では一様な公比で遅れ報告の挙動が充分再現できていると考えられるが、今後何らかの事情で突発的かつ急激な挙動の変化が発生する事象が生じた場合には、現行のモデルでは対応できない。これらの問題は、パラメータに回帰的な構造を設定し、公比が説明変数を内包する形での解決が期待される。ここで、説明変数として考えられるのは、近隣県の情報といった地理的要因や、データ収集に関する変更などである。

最後に、本モデルは罹患数の経年的変動の影響を考慮していない点も問題点であると考えられる。図１のヒートマップにおいて等高線が横向きに示された通り、愛知県における罹患数の動きは、遅れ報告に比して経年的変動の方が極めて大きな傾向にある。罹患数の経年変動は連続的と考えられるため、例えば前後の年に比べて特異に低い罹患数が報告された年が存在したならば、その年の遅れ報告は多くなることが予想される。従って、MCIJ報告の挙動の特異性を、その後の遅れ報告に反映させるような改良も必要であると考えられる。

罹患の遅れ報告については経年的なアップデートが繰り返されるが、このテーマにおける実用上の疑問点は「何年間分の遅れに関する補正が必要か」あるいは「何年分の情報を蓄積すれば信頼できる予測値が得られるか」である。この点に関して2003年罹患データを用いてベンチマークを行った結果を図９に示す。縦軸はモデルによって推定された罹患数の漸近値、横軸はMCIJからの遅れ年（＝利用したデータの数）を表す。横軸の始点を3としているのは、それ以下の1時点では初項の差分が定義できないため、そして2時点による解析では単に2時点を結ぶ結果となり意味合いが薄いと考えたからである。最も左側のプロット（3, 16065）は、横軸が3であるため、MCIJ2003、2004、2005の3年分のデータから罹患数の漸近値を推定した結果が16065人という意味である。水平方向の破線は y=15976 であり、最も信頼できる（＝利用データが最大）全てのデータ（13時点）に基づいて推定された罹患数の漸近値を表す。MCIJから3年目、4年目までのデータのみによる推定には大きなバラツキが観測されたが、5年目から急激にバラツキが小さくなり、10年目以降はほぼy=15976に近い結果が得られた。この事例に基づくと、遅れ報告に対する解析を行うには少なくとも5年先までの情報が必要であり、可能であれば10年先までのデータを蓄積されれば信頼に足る結果が得られることが期待される。ただし、この結果は2003年単年に基づく結果であり、他の罹患年では異なる結果となる可能性は否定できない。ただし、対象年の前後を含む形での解析方法が確立されれば推定に十分な情報がプールされることから、より短い観察期間でも

信頼できる結果が得られる可能性もある。

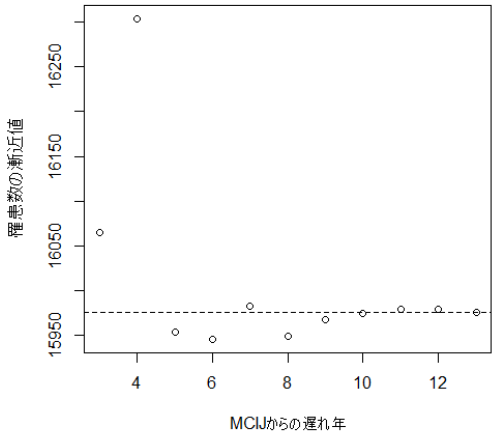

図９　2003年罹患に対するベンチマーク

罹患の遅れ報告に対する本数理モデルは、アイデアを提起した初期段階であり、様々な問題点や発展性を内包している。しかしながら、差分に対する等比数列を仮定したことにより、遅れ報告に関する補正が完了したという意味での「真の罹患数」を推定することが可能となった。

F.　健康危険情報

総括研究報告書にまとめて記入

G.　研究発表

1.　論文発表

1）　加茂憲一, 福井敬祐, 坂本亘, 伊藤ゆり. がん対策立案・評価における意思決定に寄与するマイクロシミュレーションの構築：大腸がんを事例に, 計量生物学, 41，93－115，2021.

2）　K.Kamo, K.Fukui, Y.Ito, T.Nakayama, K.Katanoda. How much can screening reduce colorectal cancer mortality in Japan? Scenario-based estimation by microsimulation, Jpn J Clin Oncol, DOI:10.1093/jjco/hyab195, 2021.

2.　学会発表

1）　加茂憲一．マイクロシミュレーションを用いた大腸がんに対する介入効果の評価：日本計算機統計学会，2021年6月4日．

2）　加茂憲一．大腸がん検診がもたらす効果のマイクロシミュレーションによる効果：がん予防学術大会，2021年9月11日．

H.　知的財産権の出願・登録状況

該当なし