資料３

"時系列解析法としてのARIMAモデルと予測"
についての概説

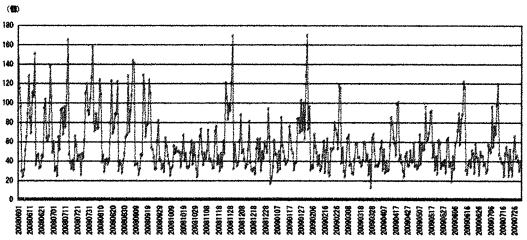
ARIMA：　Autoregressive（自己回帰）
Integrated（統合）
Moving average（移動平均）

2010年7月

特定非営利活動法人　ビュー・コミュニケーションズ

Ⅰ時系列データ　〝分析・予測の経緯〟

1　実際の時系列データの多くは、一見、不規則に激しく上下変動を伴って推移するので、将来を予測しようとするのはこれまで困難と考えられてきた。



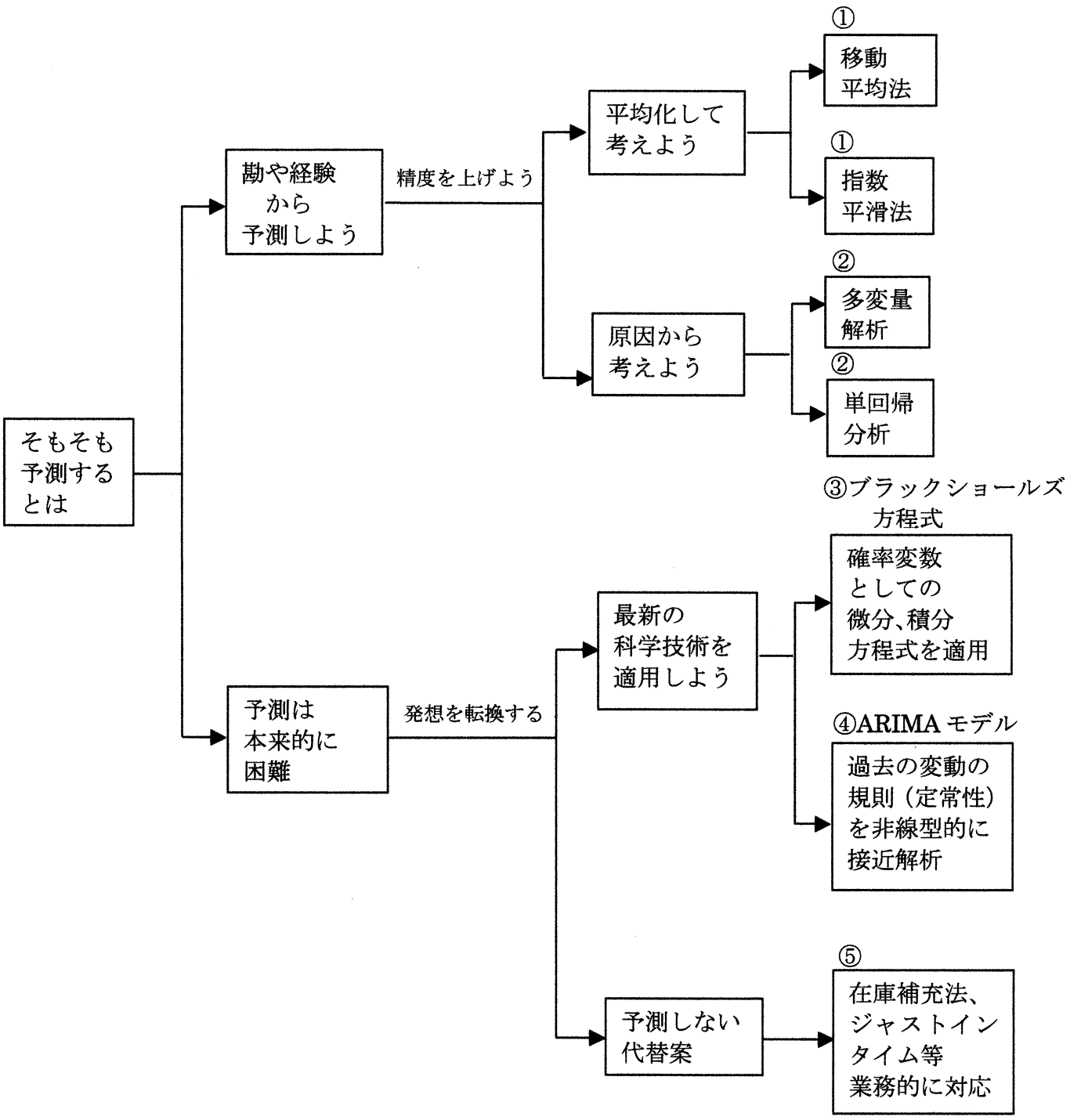
これまでの代表的な予測法①②（次頁図表　〝予測、Forecasting〟の場合分け）は、ほとんどのケースで目安や勘どころとして利用されてきたが、統計的に信頼おける結果が得られてきたとは言いがたい状況にある。

2　しかし、近年、金融工学分野でのめざましい成果（ブラックショールズ方程式の実務適用の成功）に見られるように、高度な数理統計技術が注目されてきた。

株式、債券価格のように、時間刻みがほぼ連続データと言える膨大な時系列データが存在し、微分、積分方程式（ブラックショールズ方程式）等の適用がしやすかった背景もあろう。

一方、多くの売上販売に関するデータは日、週、月、年といった離散型データで定差方程式（ARIMA モデル）の扱いとなり、一般になじみが薄く、専門家の人数も相対的に少なかったことも普及の妨げとなっていたと思われる。

〃予測、Forecasting〃 の場合分け



①変動を平均化するだけで、大きな予測誤差が生じてしまう。また、平均化、平滑化のパラメータが固定的で、変化を常とする現実に対応できない。

②過去の原因（要因）が解明できたとしても、その要因自体を予測しなければならず、いわゆる予測の予測といった矛盾が生じてしまう。
また、変化する現実を動態的に捉えるには、多変量時系列解析（VARMA）の専門分野に入り込み、実務上困難を極める。

③金融工学、グローバルな金融実務分野でここ10～20年、目覚しい成果を上げてきた。（exブラックショールズ方程式の実用化）

④定差方程式（ARIMAモデル）のパラメータ推計（最尤法）は、やや専門的で高速、大量処理が求められる実務への適用が遅れてきた。

変動を確率事象と捉えるか、定常過程と捉えるかで、数学的取り扱いがやや異なるが、基本的にはランダムウォーク※にどうアプローチするかで別れているとも言える。

※ $\frac{df(x)}{dt} = \varepsilon_t$ 、 $\varepsilon_t = (0, \sigma^2)$

3　実業、実務での様々なパターンの時系列変動データに対し、ごく少数の専門家が極めて限られた事例につき、ARIMA モデルを構築、研究してきた。実用に供するためには、高度な実証と経験が求められ、とても現場において常時使用できる状況にはなかった。

当法人では、ここ数年 ARIMA モデルの解析力、現場適用性ついて研究を進め、一見するとパターンが見られない変動を示すことが日常的に生ずる現実の時系列データ（2000 を超える種類：種々の日次データ、週次データ、月次データ、年次データ、マクロデータ、ミクロデータ、GDP、市場全体、人口、食品売上高、商品販売高、個人販売高 etc.）でその適合性を検証してきた。その結果、ARIMA モデルは現実データに対する説明力、予測力で予想を越えた極めて優秀な適合性を持つことが判明した。

例えば、百貨店における実証研究結果（950 ブランドの週次データ　それぞれに ARIMA モデルを適用し予測を実施、2010 年）では、予測値と実績値の差（誤差）は次のようになった。

誤差　（週次データ、31 週平均モデル誤差）

1%　未満	448 ブランド	(47.2%)
1～3%	329 ブランド	(34.6%)
3～5%	89 ブランド	(9.4%)
5～10%	62 ブランド	(6.5%)
10～26%	22 ブランド	(2.3%)
	950 ブランド	(100%)

$$\blacksquare \quad \frac{\text{31 週モデル推計値} - \text{31 週実際値}}{\text{31 週実際値}} \times 100\%$$

4　時系列解析の系譜

～1960 年代

・経験的方法の発展（移動平均、指数平滑化法、季節調整法 etc.）

・純理論的方法の発展（周波数領域分析の数学的理論 etc.）

1970 年代

・Box = Jenkins　(1970,1976)

〃Time Series Analysis : Forecasting and Control〃 (Holden－Day)

Box＝Jenkins 法の提唱（一変量時系列分析）

・赤池・中川 (1972)

「ダイナミックシステムの統計的解析と制御」(サイエンス社)

AIC（赤池情報量基準）の開発

1980 年代

・理論的発展（多変量、非正規性、非定常性、非線形性などについての統計的扱いの精緻化 etc.）

・応用分野の拡大（工学・理学・医学・マクロ経済学 etc.）

1990 年代　？

コメント：日本の研究水準

時系列分析の分野は統計学・計量経済学において最も国際競争力のある分野であり、幾つかの問題では世界の学会をリードしている独創的研究もある。

しかし、理論の各分野への本格的応用は始まったばかりである。

(H5. 国友、小松　共同セミナーより抜粋)

5　時系列解析の定番的文献

(計量経済学分野の大学院で広く読まれている教科書例)

山本　拓	経済の時系列分析　1988
James Hamilton	Time Series Analysis　1994
P.J.Brockwell and R.A.Davis	Introduction to Time Series and Forecasting　2000

6　適用領域の可能性

ARIMA モデルのような、汎用性が高い時系列解析技術は、下図の領域に幅広く適用されると期待される。

分野 / 領域	行　政	産　業	科　学	教　育	・・・・
計　　画					
管　　理					
戦略・戦術					
評　　価					

参考　～　多変量解析を予測に用いようとするケース

小売チェーン店X社（100店舗）が、商品Aの販売予測を多変量解析を用いて行う場合を考えてみる。

（個別モデル構築をあきらめ、全体モデルで考えようとする）

まずは、個々の店舗における商品Aを多変量解析で予測しようとすると、全体で100モデル構築しなければならず、煩雑であるので、100店舗合算して考えようということになりがちである。

よくありがちな多変量解析モデルの例（100店舗合算モデル）

商品Aの売上＝来店客数×商品Aの来店客購買率

商品Aの来店客購買率は一定と仮定し、来店客数が

商品Aの価格指数（ex価格弾力性値）、気温、天候、商圏人口、競合性指標に依存するとして多変量解析をするような場合である。

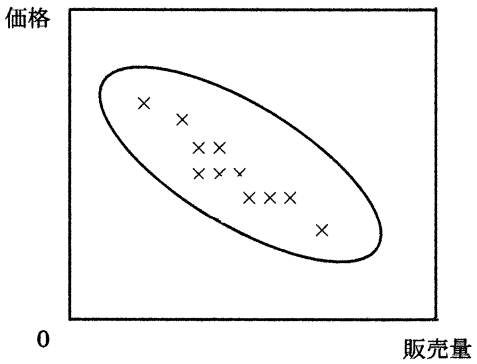
（　来店客数＝a_1×価格指数＋b_1×天候指数＋c_1×商圏指数＋A　）

⇒ 現実には、地域の諸事情は異なるので商品Aの売れ行きは各店舗でかなり異なっている。

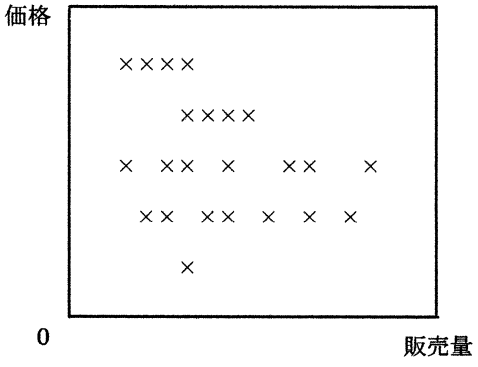
（全体では良さそうでも、個々に見ると大きく異なる）

100店舗合計した場合の商品Aの価格と販売量の関係図が、図表－1　のように単相関しているように見えるので、来店客数の説明変数に組み込もうとする。

（図表－1）100店全体



（図表－2）個別店舗

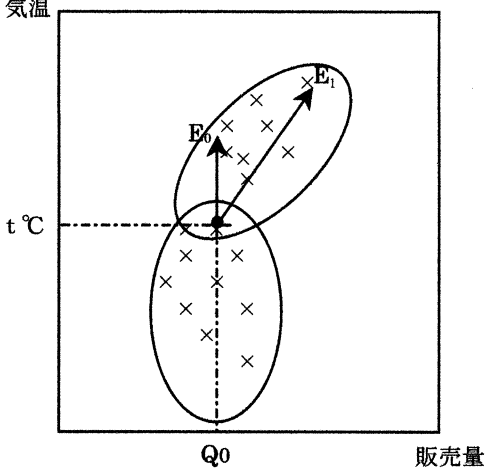


しかし、実際に個店における価格と販売量の関係（100～150日）を調べてみると図表－1のように同一価格に対する販売量は大きくバラけており、単相関を仮定するのはムリ（統計的には無相関）である。
したがって、例えば特売価格設定をこのような多変量モデルで決定しようとすると、個店ベースでは販売量予測値と実際の値とが大きくズレてしまうことが頻発する。

（変化を追えない固定的モデルとなっている）

気温、天候についても同様の問題があるが、何よりも、気温、天候の変動パターンが変化しやすいので、固定化された相関モデル（多変量解析モデル）では、現実を説明するのは困難になってしまう。

（図表－3）気温と販売量の関係



例えば、ある商品の売れ行きが気温t℃を境に、気温上昇とともに大きく変わると考えられているとしよう。（図表－3）
明日、明後日と気温がt℃を超えると予想されているので、発注担当者は大きく発注量を増やして、対策をとったとしよう。しかし、この発注担当者の実際の経験からは、このような考えに基づく発注の正解・不正解は運・不運によると思っているはずである。

図表にそって考えてみる。
明日、明後日の気温がt℃をどの程度上回るのか、天気予報の精度は必ずしも高くないこと、気温上昇の精度にもよろうが、図表中のE_0になるかE_1になるか、その確率は不明であること等で、気温上昇による販売量増がどの程度実現するか、極めて不確かな事と言えよう。
一般的に言えば、気温、天候の変動パターンは変化しやすく、それに伴う人々の適応行動も変わりうるので、時間の流れにおけるダイナミックな関係がわからないと、発注対策は単なる運、不運の世界に落ち込んでしまう。

ストアオペレーション上、常に精度の高い気温、天候予測を入手することは現実的に難しく、仮に入手できたとしても、固定的なモデルでは、人々の適応行動としての買物行動を捉えきれず、満足できる予測結果を得られないこととなる。

（変化し続ける商圏人口、競合関係）

首都圏の都市部で、住民基本台帳にもとづいて、中学校区（or 小学校区）単位で捉えた年齢階層別流入、流出人口を調べると、毎月少しづつ変化し、年々で増加を示す学区と減少または横バイを示す学区にわかれる。さらに増加・減少傾向についても様々なパターンにわかれ、人口移動が日々ダイナミックに起こっていることがわかる。

しかし、各店舗別の対象商圏を支えるこのような月々の人口移動データは把握されておらず、出店時に調査されたであろう推計商圏人口をほぼ固定した形で日々のストアオペレーションがなされているのが実情であろう。

つまり、多変量モデル構築時に設定した商圏人口を固定したまま、各種オペレーション戦略、戦術が実行されていることとなる。

人口データは公的に把握されているのでまだしも、競合関係となると、商圏内競合店の動態を常時把握するのは技術的にも経済的にも極めて困難性が高いので、例えば継起的に実施される競合店価格調査程度に限られ、多変量モデルの予測精度はほとんど保証されないものとなっている。

（多変量解析モデルの適合性について）

多変量モデルは上記の様々な理由から、予測モデルとして活用しようとするのは、大いにムリがあるが、過去に存在している売上を支える諸要因を探るという意味では有用である。また、例えばウィルス感染数を予測しようとするのではなく、ウィルス感染を防ぐための対策を立てる意味　　で、多変量モデルにより種々の感染ルートを解析し防護等に役立てるには有効である。

Ⅱ ARIMA モデルの考え方

1　ARIMA モデルの概念

(1) 原因からおわずに、結果からおう

過去の自分自身の足跡（販売量、生産量等の変動量）は、近い将来の足跡を占う

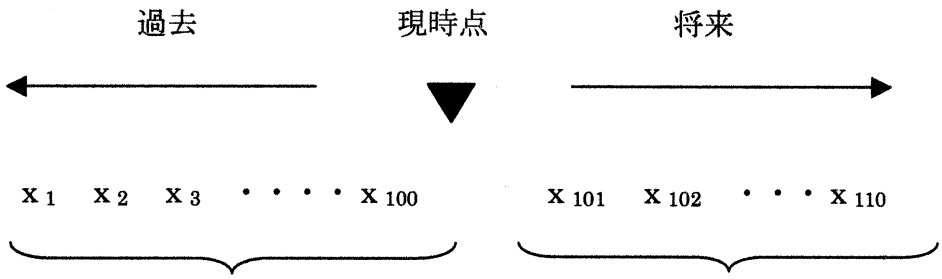
‖

過去に存在していた足跡の規則（定常性）を数理統計（ARIMA モデル）を用いて解析

‖

過去のすべての時点の足跡の間で、どのような相互関係があったか（自己回帰性）解析し、その関係性が短い時間は継続すると考え、予測値として扱う。

(2) 例えば 100 日間の販売データが与えられたとする。



100 日分の過去の販売データ　　10 日分の販売予測データ

ARIMA モデルは、

過去の実績データ $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_{100} \end{bmatrix}$ を用いて、近い将来の値 $\begin{bmatrix} x_{101} \\ x_{102} \\ \cdot \\ \cdot \\ \cdot \\ x_{110} \end{bmatrix}$ を予測する

ので、販売データの背後に存在する様々な構造（ex 特売パターン、天候変化、競合店の値引戦略、客層の変化・・・・）は全く不知として解析している。

(3) ARIMA モデルとして使用する過去のデータ量については、おおむね 30～40 データ（例えば日、週、月次データどれでも）以上あればモデリング、予測が可能となる。
理論上しばしばデータ量は多ければ多いほど望ましいとされるが、実務的、現実的に見て限界がある。
つまり、「遠い過去のデータ変動規則を直近に反映させるのは望ましくない」との考えるのが妥当で、使用するデータ量はせいぜい 200～300 データ当りが最大であろう。
例えて考えてみる。
仮に江戸時代まで GDP が正確にわかっていたとしても、200 年、300 年の日本の年次 GDP データを使って、来年以降 3～5 年の GDP を予測しようとするのは無理がある。江戸時代の GDP と 2010 年の GDP に有意な相互関係があると考えるのは変であるからである。
同様に、365 日前のデータを用いて、向こう 2～3 日間の豆腐の販売予測しようとするのも無理が生じる。特殊な日（ex バレンタインデー、お正月・・・）を除き、昨年の同日の販売量（購買量）と今年のその日の販売量との相互関係性は薄く、最近の販売事情、天候、家族内諸事情等に依存して販売量が決まってくると考えるのが自然である。

(4) ARIMA モデルは短期予測に用いられ、長期予測には不向きである。過去データ量を N とすると、予測適量は 0.1～0.2N が妥当である。
この考え方は、後述する ARIMA 逐次更新モデル（動的モデルの構築）とも関係する。
過去の期間（データ量で N）になりたっていた関係性（定常性）は、将来

になればなる程、その関係性そのものが変化していると考えられるので、
0.1～0.2N 程度であれば関係性継続の確率が高いと想定しているのである。

2　ARIMA モデルの数学的表現

ARIMA モデルとは、自己回帰和分移動平均モデルで次の定差方程式で表される。

$\phi(B)X_t = \theta(B)\varepsilon_t$
$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$
$BY_t = Y_{t-1}$　(B：ラグ演算子)
ϕ_i　$(i = 1,2,\cdots,p)$　　自己回帰パラメータ
θ_j　$(j = 1,2,\cdots,q)$　　移動平均パラメータ
ε_t　は誤差項であり、$\varepsilon_t \sim (0,\sigma^2)$で定義されるホワイトノイズ

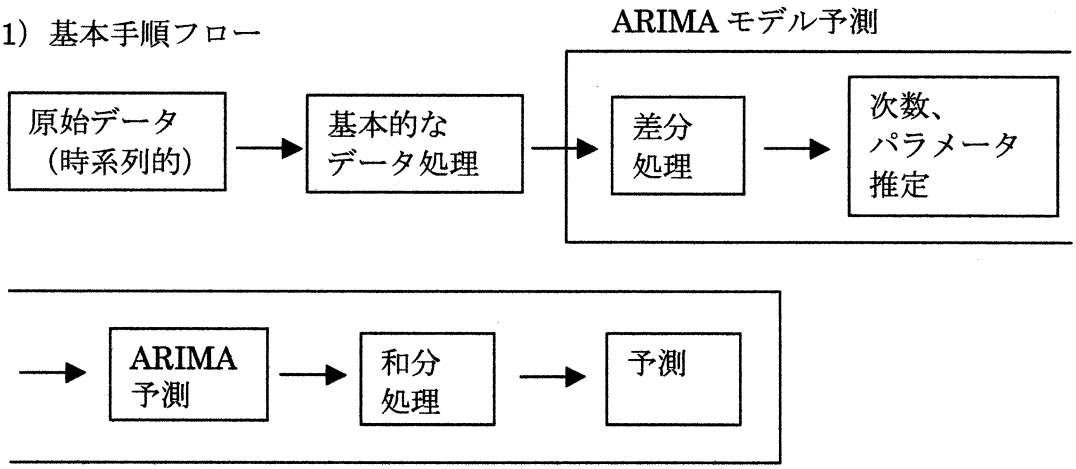
与えられた実際の時系列データ $x_t = \{x_1, x_2, \cdots, x_t\}$ に対して、パラメータ (ϕ_i, θ_j) や分散（σ^2）を推定する問題を解くのがＡＲＩＭＡモデルと言える。

推定の基本骨子は、情報量基準（ＡＩＣ等）を用いて、自己回帰次数、移動平均次数 p、q を推定すること、及び、実際値と予測値の差（誤差）が最小となるパラメータ ϕ_i、θ_j を最尤法（Maximum Likelihood Method）を用いて推定することにある。

ＡＩＣは赤池教授によって提唱されたモデルの当てはまり性に関する基準で、複雑性と説明力のトレードオフ関係に注目して定めたものである。最尤法は、標本が観測される確率（尤度）を最大化する方法で、対数尤度関数の偏微分＝０から各種繰り返し接近法により、パラメータ推定を行う高度な技法である。
自己相関、偏自己相関、ＡＩＣ、ＳＢＩＣ、最尤法、残差検定等の数式は専門的になるので、ここでは省略する。
（詳細は参考文献を参照されたし）

3　ARIMA モデル構築及び予測の手順

(1) 基本手順フロー



(2) 基本的なデータ処理　～　何について分析・予測するのか

原始データ（99 日）

年月日	商品名	発注入数	納品数量	売上金額	売上数量	廃棄金額	廃棄数量	値引金額	値引数量	…
20100201	天然水２０００ｍｌ	6	18	4	0			0		
20100202	天然水２０００ｍｌ	6	6	9	0			-1		
20100203	天然水２０００ｍｌ	6	6	7	0			0		
20100204	天然水２０００ｍｌ	6	6	12	0			0		
20100205	天然水２０００ｍｌ	6	12	11	0			-2		
20100206	天然水２０００ｍｌ	6	12	46	0			-2		
20100207	天然水２０００ｍｌ	6	750	738	0			-184		
20100208	天然水２０００ｍｌ	6	24	7	0			0		
20100209	天然水２０００ｍｌ	6	30	5	0			0		
20100210	天然水２０００ｍｌ	6	6	5	0			0		
20100211	天然水２０００ｍｌ	6	6	22	0			-1		
20100212	天然水２０００ｍｌ	6	18	14	0			0		
20100213	天然水２０００ｍｌ	6	1212	301	0			-77		
20100214	天然水２０００ｍｌ	6	0	351	0			-47		
20100215	天然水２０００ｍｌ	6	0	6	0			-6		
20100216	天然水２０００ｍｌ	6	0	6	0			0		
20100217	天然水２０００ｍｌ	6	0	18	0			0		
20100218	天然水２０００ｍｌ	6	0	8	0			0		
20100219	天然水２０００ｍｌ	6	0	17	0			0		
20100220	天然水２０００ｍｌ	6	0	55	0			0		
20100221	天然水２０００ｍｌ	6	600	769	0			-660		
20100222	天然水２０００ｍｌ	6	0	9	0			0		
20100223	天然水２０００ｍｌ	6	0	16	0			0		
20100224	天然水２０００ｍｌ	6	0	13	0			0		
			∫							
20100420	天然水２０００ｍｌ	6	0	19	0			0		
20100421	天然水２０００ｍｌ	6	0	25	0			0		
20100422	天然水２０００ｍｌ	6	0	11	0			0		
20100423	天然水２０００ｍｌ	6	0	10	0			0		
20100424	天然水２０００ｍｌ	6	0	31	0			0		
20100425	天然水２０００ｍｌ	6	480	351	0			-7		
20100426	天然水２０００ｍｌ	6	0	13	0			0		
20100427	天然水２０００ｍｌ	6	0	19	0			0		
20100428	天然水２０００ｍｌ	6	0	14	0			0		
20100429	天然水２０００ｍｌ	6	0	56	0			0		
20100430	天然水２０００ｍｌ	6	0	21	0			0		
20100501	天然水２０００ｍｌ	6	0	13	0			0		
20100502	天然水２０００ｍｌ	6	480	503	0			0		
20100503	天然水２０００ｍｌ	6	0	19	0			0		
20100504	天然水２０００ｍｌ	6	0	16	0			0		
20100505	天然水２０００ｍｌ	6	0	32	0			0		
20100506	天然水２０００ｍｌ	6	0	21	0			0		
20100507	天然水２０００ｍｌ	6	0	15	0			0		
20100508	天然水２０００ｍｌ	6	0	47	0			0		
20100509	天然水２０００ｍｌ	6	780	545	0			0		
20100510	天然水２０００ｍｌ	6	0	11	0			0		

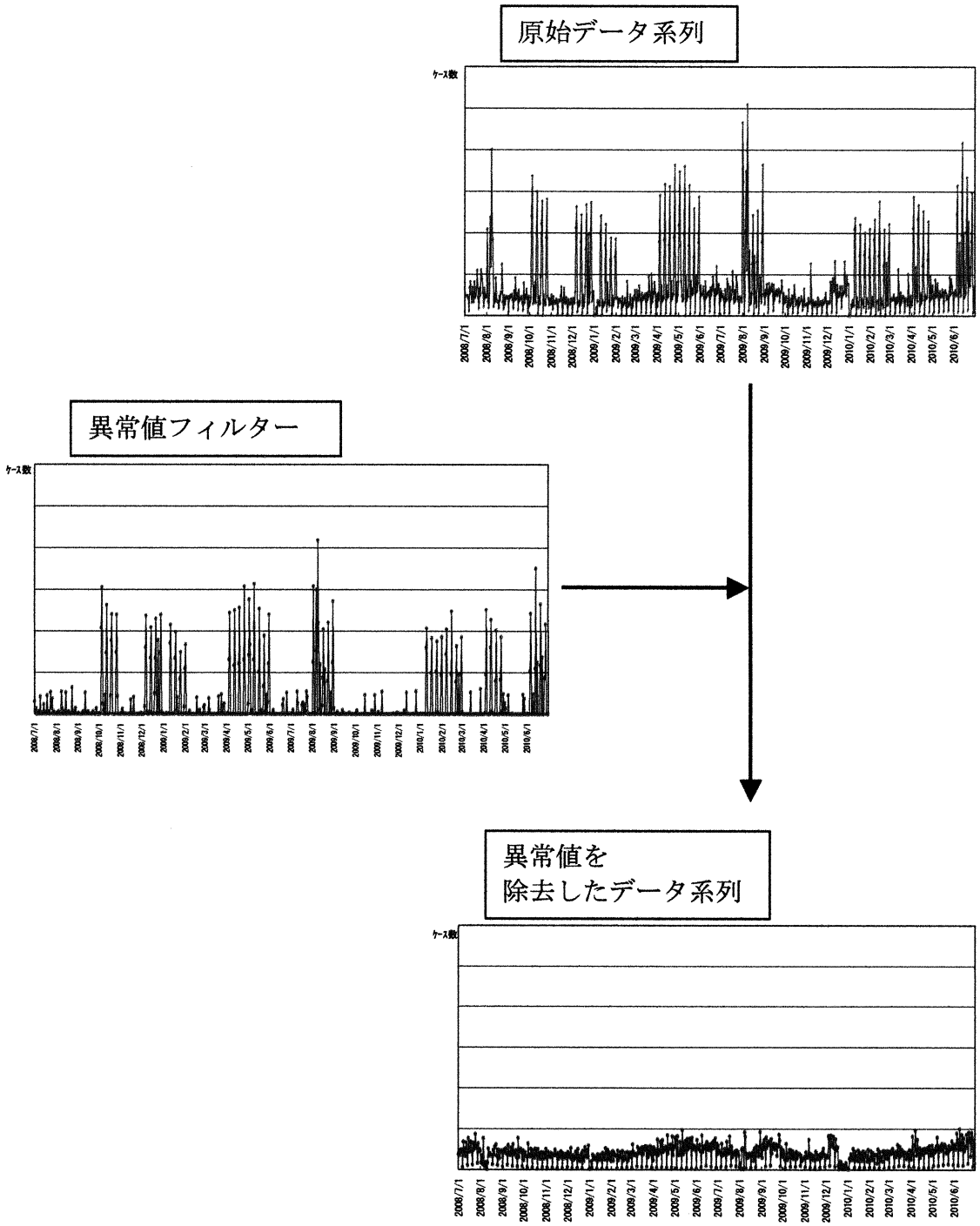
（基本的な処理）

- ☑ 誤ったデータの修正
- ☑ 日付等の連続性確保
- ☑ 欠損値処理
  （0 とおく、分析対象外とする etc）
- ☑ 異常値を除去する* or しない
- ☑ データ系列が指数的遷移を示す場合には対数変換などを施す

⇩

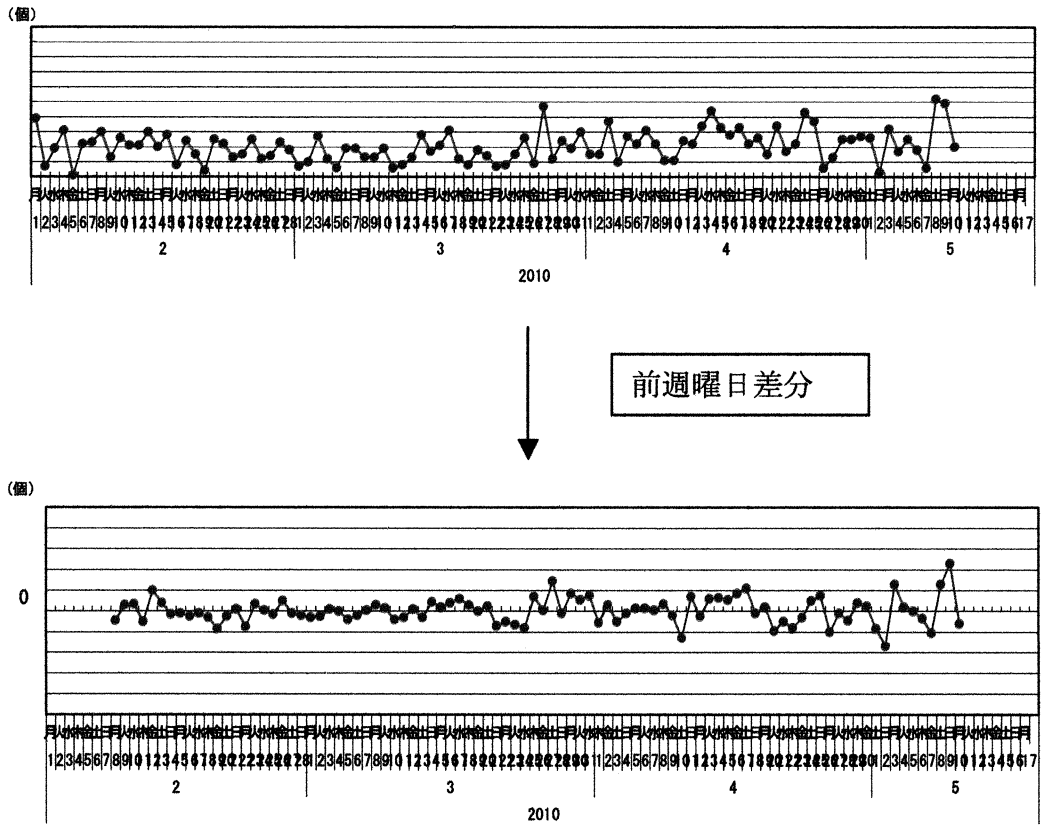
ARIMA 分析するデータ系列を抽出

＊異常値除去の例



異常値フィルターは様々にあるが、スーパー等のケースでは特売フィルターがその例である。

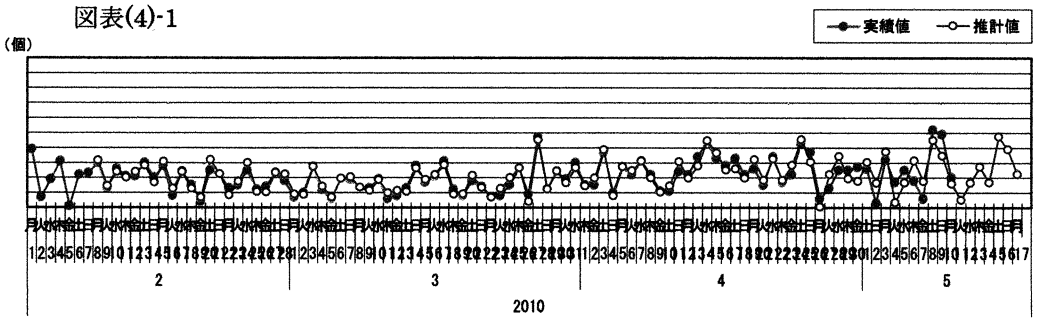
(3) 差分処理



もとのデータ系列には、トレンド（長期趨勢的傾向）や季節性・循環性（曜日、季節などの販売特性、買い替えサイクル・・・）が含まれている。
この部分を取り除くために差分処理を施す。
前年同月、前週同曜日、前日などの差分があり、どの差分が良いか事前には良くわからないことも多いので、様々な差分につきモデリング推計作業を行うこととなる。

(4) ARIMAモデルの中核的な部分

図表(4)-1



差分処理されたデータ系列につき、定常性（規則性）がありそうと考え、ARIMAモデルを用いて各時点における数値間の関係性を解析する。自己相関、偏自己相関、情報量基準などの指標をもとに、AR（自己回帰）次数、MA（移動平均）次数　～　ある時点のデータ数値とどれくらい前の時点のデータ数値とが関係しているか　～　を設定し、非線形の接近法（最尤法）によりAR、MAの各次数に対応するパラメータ値を推定する。このようにして構築されたモデルにつき、残差検定を行い、統計的信頼性があると判定されたものにつきモデリングが完了する。図表(4)-1は、そのモデルから求められる推計値と実際値との関係をグラフ化したものである。

ex 実際に推計されたモデル数式（ARMA（2，2）の例）

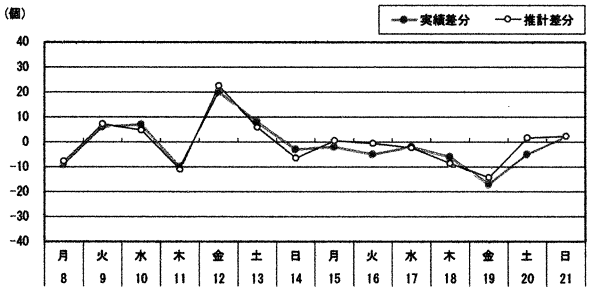
$$X_t = 0.2633\,X_{t-1} + 0.8979\,X_{t-2} + Z_t - 0.07593\,Z_{t-1} - 0.9997\,Z_{t-2}$$

（Z_t = ホワイトノイズ　σ_Z（分散）$= 0.12759 \times 10^3$）

この例は、スーパーマーケットにおける〝ぶた挽肉〟の日々（t）の販売数量（Xt）のモデルであり、観測期間（このケースでは92日間）を通じて、前日及び前前日の販売量と当日の販売量とが関係し,このような式が成り立っていた。

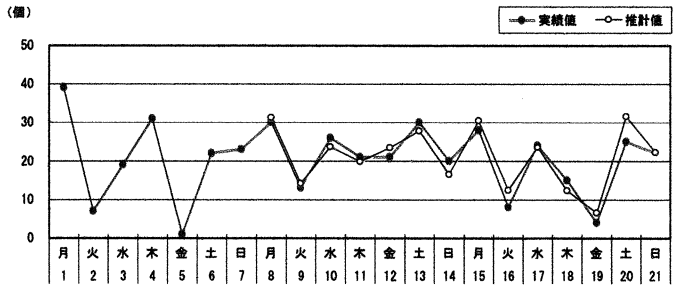
この事例における実際値とモデル推計値（差分系列）

期間	実際値	推計値
2／8（月）	−9	−8
9（火）	6	6.9
10(水)	7	4.5
11(木)	−10	−11.4
12(金)	20	22.2
13(土)	8	5.6
14(日)	−3	−6.7
15(月)	−2	−0.2
16(火)	−5	−0.8
17(水)	−2	−2.6
18(木)	−6	−8.9
19(金)	−17	−14.7
20(土)	−5	1.3
21(日)	2	2
	・	・
	・	・
	・	・



差分化されたデータ系列に対する ARIMA モデル構築が終わると、次にもとのデータ系列への対応をとるために、和分過程（ AR①MA 、Integrated ）により、ARIMA 推計値が与えられることとなる。

和分過程は、（1 週間前の同曜日の販売量＋当週の同曜日の差分推計量）である。



このグラフは、図表(4)-1　の該当期間を拡大して示してある。

4　ARIMA 予測と異常値,外部干渉(衝撃)

(1) ARIMA 予測における異常値の扱い方

過去の時系列データセットに異常値、不確定値があった場合でも、基本的に,そのままのデータで（データの真実性が与えられている限り)、ARIMA モデルで解析することとなる。

例えば、ある日の異常値（ex 通常よりも明らかに異常に売れた）は,モデル上それが生じたことを事実として捉え、その後、そのような値が発生するはずと考え、予測値を出すこととなる。

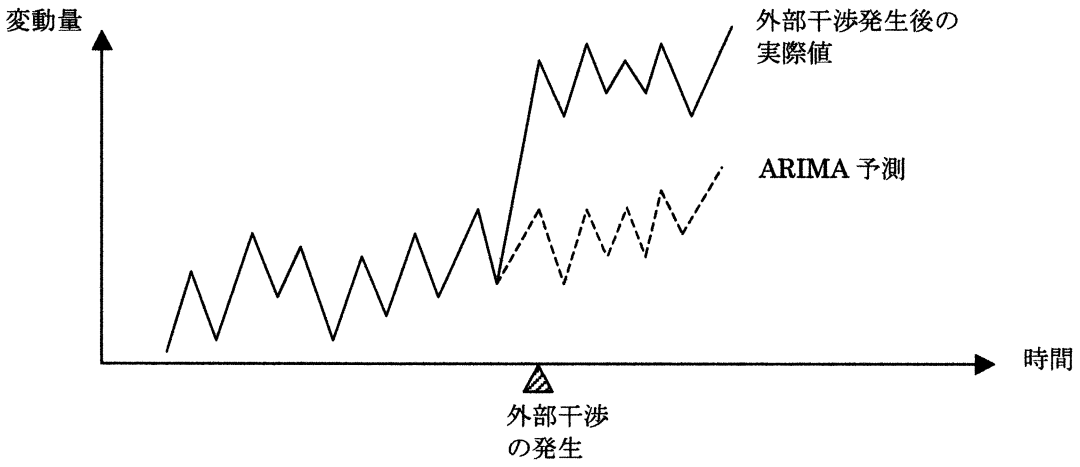
異常値の発生は事前には予測できないので、発生した時点では実際値＞予測値となり大きく乖離する。

その後は、異常値が発生したと事実認識し、その後を予測するので、予測値＜実際値となり､先の予測誤差が相殺されて行くこととなる。いわば自動的に補正されて行くことになるので、一過性的異常値は、特別な処理をしないで、ARIMA モデルに任せておいた方が良いこととなる。

(2) 大きな外部干渉(衝撃)

大きな自然災害、大きな制度の変更（ ex 消費税、年金)、画期的な新技術､新製品の登場など、過去の定常過程になかった事象が発生する場合がある（時系列解析の分野では外部干渉（衝撃）と呼ぶ)。

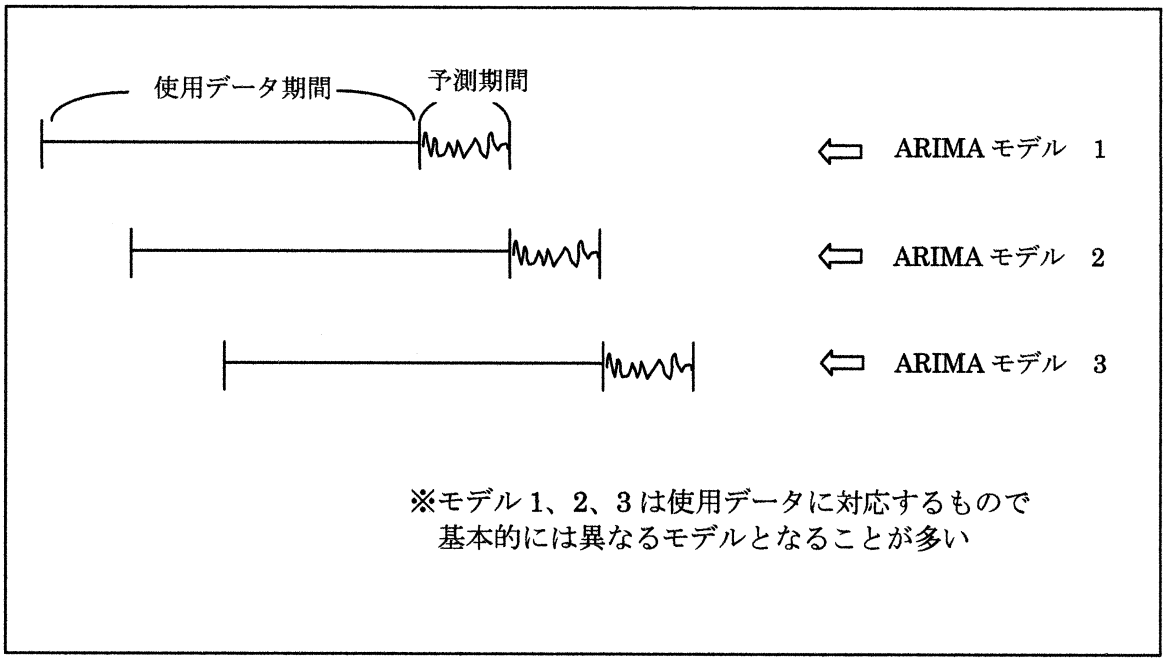
ARIMA モデルでは、勿論このような事象を予想することはなく、実際値と予測値が大きく乖離することとなる。



外部干渉の大きさによるが、一般には、外部干渉発生後の短い期間の実際値の変動パターンをそれまでの ARIMA モデル（一定の規則をもった変動パターン）に合成することで、その後の推計・予測を行う方法をとることとなる（伝達関数モデル）。

5　動的なモデルへ　～　逐次更新モデルの適用

(1) 時々刻々変化する経済変動をトレースするために、静態的な ARIMA モデルでは不十分で、モデルを逐次更新型で再構築しつづけることが必要となる。逐次更新パターンは、週次更新、4 半期更新など様々で、それぞれのビジネスサイクル等に応じて設定される。



(2) 景気転換点と逐次更新モデル

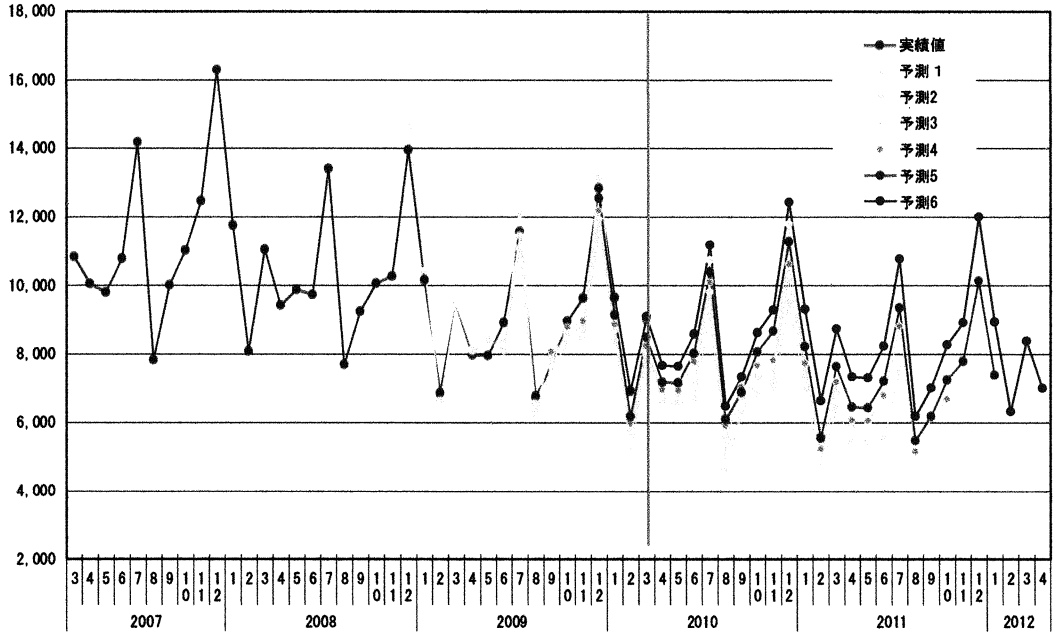
静態的 ARIMA モデルに用いられる差分法（和分法）は、中期的な景気サイクルについて十分対応できない。

しかし、逐次更新モデルを適用すると、次の例のように、短い時間で景気転換に予測がキャッチアップすることとなる。

（百貨店における売上転換点の事例）

逐次更新モデル

予測1	ARIMA		(2、D12D1、3)
予測2	ARIMA	log	(2、D12D1、0)
予測3	ARIMA	log	(2、D12D1、0)
予測4	ARIMA	log	(0、D12D1、1)
予測5	ARIMA	log	(0、D12D1、1)
予測6	ARIMA	√	(0、D12D1、1)



2009年9月～12月　を転換点とした売上トレンドの変化が見られる。

予測 1,2,3 では、ほぼ一様に売上高のダウントレンドを予測したが、

予測 4 を境に、予測 5, 6 と反転して上昇トレンドを予測している。