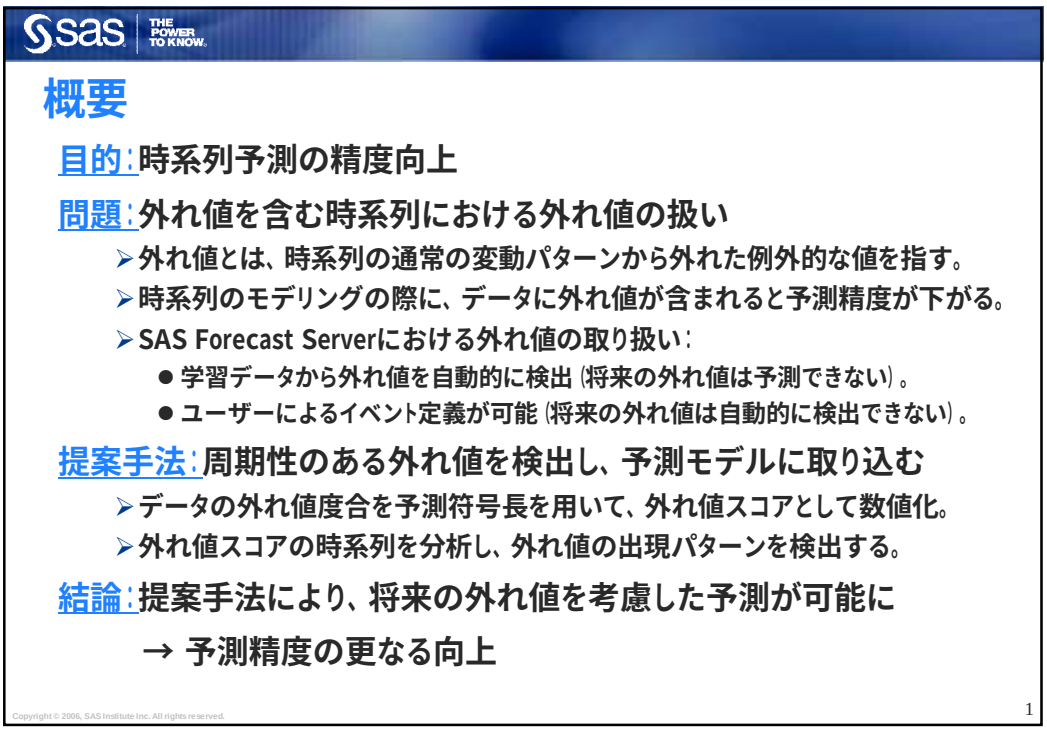


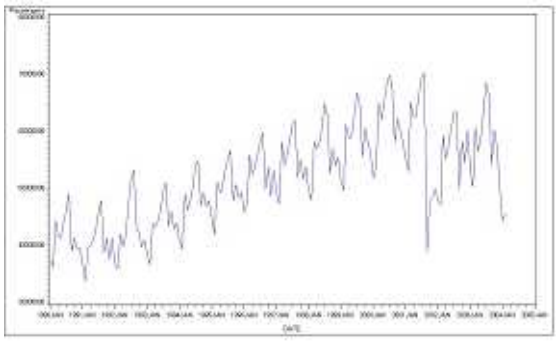
時系列データの予測

目的：時系列データの将来の値を予測すること

入力データ

- 一次元の数値時系列（複数あってもよい）
  例. アメリカの飛行機利用者人数の推移

縦軸；人数



横軸；日時（1990年1月から2004年6月まで）

時系列データの予測の重要性

リスク管理の必要性

- 業務において扱う多くの量が時間的に安定せず変動している。
  - 商品の在庫、電力の需要、など。
- 変動への対処を怠ると多大なリスクが発生する可能性がある。

変動の予測

- リスク管理の観点から、将来の変動を予測し、その結果を基に事前に対策を講じることが重要となる。
- 予測の精度が高いほど、リスク管理は容易になる。
- 多大なデータから人手で予測することは、コスト的にも時間的にも困難。

時系列モデリングによる予測

- SAS Forecast Server(以下「FS」)を用いて高精度な予測を大量に速く自動的に実行可能

外れ値を含む時系列の予測

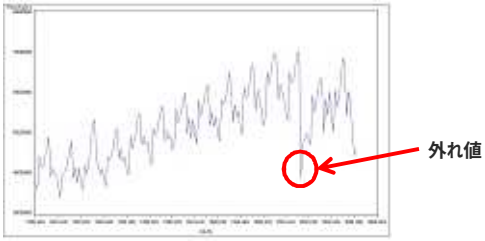
本発表では、外れ値を含む時系列の予測について取り扱う。

問題設定

目的：時系列予測の精度の向上

入力：外れ値を含む時系列データ

- 外れ値とは、時系列の通常の変動パターンから外れた例外的な値を指す。



解くべき問題

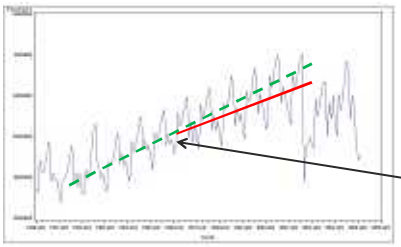
- 出来る限り高精度に、外れ値を含む時系列の将来の値を予測すること。

4

外れ値を含む時系列予測への要請

外れ値の検出

- 学習データに含まれる外れ値の自動検出
  - 外れ値を考慮せずに予測すると、外れ値の影響で予測精度が下がる。
  - ユーザーが外れ値か否かを目視で判断するのは困難なことが多い。



将来の外れ値の予測

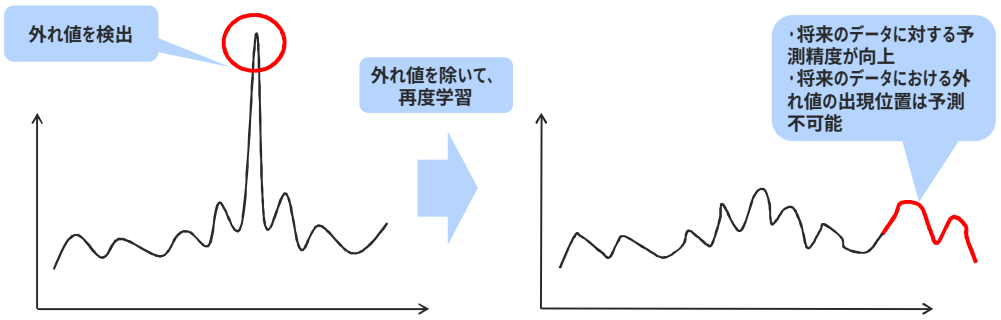
- 将来のデータにおける外れ値の出現位置を予測する必要がある
  - 将来の予測における外れ値の影響も考慮することで予測精度を更に向上させたい。

5

要請を満たすための手法：現状と課題(1/3)

既存手法①：時系列モデリングによる外れ値の自動検出

- 推定されたモデルから外れている点を外れ値として検出→外れ値を除外して再度モデル推定→推定されたモデルから新たに外れ値を検出→…
- モデルは自動的に推定されるので、外れ値の検出も自動的。
- 外れ値のモデル化をするわけではないので、将来のデータにおける外れ値の出現位置の予測はできない。

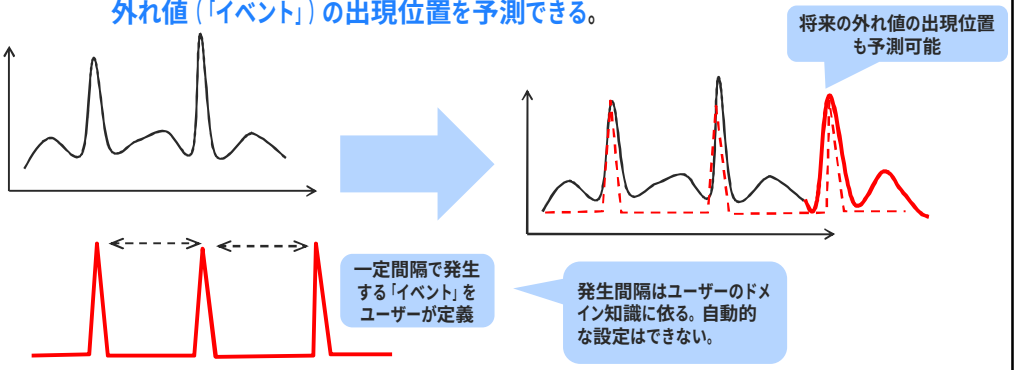


6

要請を満たすための手法：現状と課題(2/3)

既存手法②：イベントの定義と挿入

- ユーザーの事前知識を基に、外れ値が発生するであろう時点（クリスマス、忘年会、土用の丑の日、等）を「イベント」として定義し、予測に取り込む。
- ユーザーのドメイン知識に依るので、自動的な検出は出来ない。
- いつイベントが発生するかは判っているので、将来のデータにおける外れ値（「イベント」）の出現位置を予測できる。



7

前頁の要請を満たすための手法：現状と課題(3/3)

課題

過去の外れ値は検出できるが、将来の外れ値の発生時期を自動的には検出できない。

SAS Forecast Server

モデリングによる時系列予測

- 時系列の時間遷移をモデル化。
- 推定したモデルに従って将来の値を予測。

高精度な予測

- 様々なタイプのモデルが用意され、その中から最適なモデルが選択される。
- トレンド、周期性（季節変動）、等、業務上必ず発生する変動パターンを考慮。
- イベントの定義と外れ値検出の機能を有し、外れ値を考慮した予測が可能。

Forecast Serverでできることとできないこと

- （可）外れ値の自動検出（ARIMAモデル）
- （可）イベントの定義と挿入
- （不可）将来の外れ値の出現位置の予測

→本発表ではこの問題を解決し、予測精度の向上を狙う。

関連研究

Forecast Serverにおける外れ値検出

- SAS/High Performance Forecasting User's Guide.

オンラインの外れ値検出

- K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD2000)*, 2000.

→確率モデルを用いた外れ値度合いの数値化をオンラインで実行する。
予測には言及しておらず、あくまで外れ値検出が主題。

予測符号化の符号長でモデルを評価

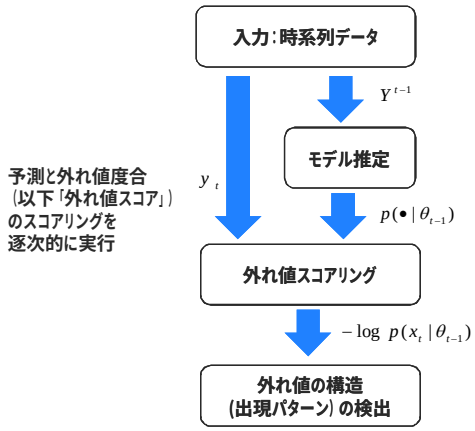
- J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transaction on Information Theory*, 30:629-636, 1984.

→予測符号化の符号長が最小となるモデルを最適なモデルとみなす。

提案手法：外れ値構造の検出を用いた予測

提案手法の概要

- SAS時系列予測プロシジャを用いた、時系列の外れ値構造の検出



予測と外れ値度合
(以下「外れ値スコア」)
のスコアリングを
逐次的に実行

「テストデータへの当てはめ統計量」
(予測符号長) を用いて外れ値を数値化

外れ値スコア時系列の構造を分析して
予測に活用

なぜ提案手法を用いるのか

将来のデータへの予測精度の向上

- 外れ値を数値化し、そこから外れ値の構造を予測するので、個々の外れ値にひとつの変数を割り当てる必要がなくなり、モデルの複雑度が減少する。
- 将来の外れ値の出現位置をモデルに取り込めるので、予測力が向上する。

工数の削減

- 人手によるイベントの検出を（半）自動化することで、工数を削減できる。

可読性の向上

- 外れ値度合（外れ値スコア）の数値化により、外れ値（イベント）の発生パターンを目で見て確認できる。

ステップ1：確率モデリング

時系列の変動を表す確率モデルを推定

- モデルの例： ARIMAモデル（詳細についてはAPPENDIX参照）

ARIMAモデルでは以下の三つの成分によって時系列の時間発展をモデル化する。

1. 平均値周りの振動（上がったら下がり、下がったら上がって平均の周りをうろうろする）成分
2. トレンド（平均値が上昇または下降する）成分
3. イベントの発生など、短期的にしか影響を及ぼさない突発的な成分

推定したモデルを用いて新規データの出現確率を算出

時刻$t-1$までのデータ：Y^{t-1}

Y^{t-1}から推定された確率密度関数: p_{t-1}

密度関数p_{t-1}のパラメータ：θ_{t-1}

時刻tのデータ：y_t

過去のモデルから見た現在のデータの出現確率：$p_{t-1}(y_t \mid \theta_{t-1})$

ステップ2：外れ値スコアリング

新規データ点の外れ値度合いを数値化

- 新規データの外れ値度合を外れ値スコアとして数値化する。
- 数値化によって、元の時系列から外れ値スコアの時系列を得る。

外れ値スコアを「テストデータへの当てはめ統計量」として定義

- 過去データを用いた予測からのずれ（予測符号長）として外れ値スコアを定義。

$$\text{外れ値スコア} \quad (y_t) \equiv -\log\ p_{t-1}(y_t \mid \theta_{t-1})$$

- 数値化によって、元の時系列から外れ値スコアの時系列を得る。
- 外れ値スコアはモデルの推定によって得られる過去のデータの出現パターンから新規データがどれだけ外れているかを表す。値が大きいほど過去の傾向から外れている、つまり外れ値らしい、ということになる。

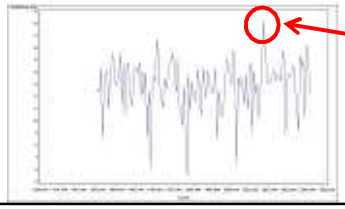
※K. Yamanishi, J. Takeuchi, et al. KDD2000, 2000.
※J. Rissanen. *IEEE Transaction on Information Theory*, 30:629-636, 1984.

14

ステップ3：外れ値スコア時系列の構造の検出(1/4)

外れ値スコア時系列の構造：

- 非周期的なピーク：モデル推定の際に考慮すべき、学習データに含まれる突発的なイベントに相当。
- **周期的なピーク：モデル推定と時系列予測の両方で考慮すべき定期的なイベントに相当。**
- 上昇トレンド：時系列の変動の仕方がFSに用意されているモデル族ではうまく表現できなくなってきているという傾向を表す。
- 下降トレンド：時系列の変動の仕方がFSに用意されているモデル族でうまく表現できるようになってきているという傾向を表す。
- 周期変動：真のモデル自体が周期変動している可能性を示唆する。



：外れ値スコア時系列。横軸は時間
縦軸は外れ値スコアを表す。

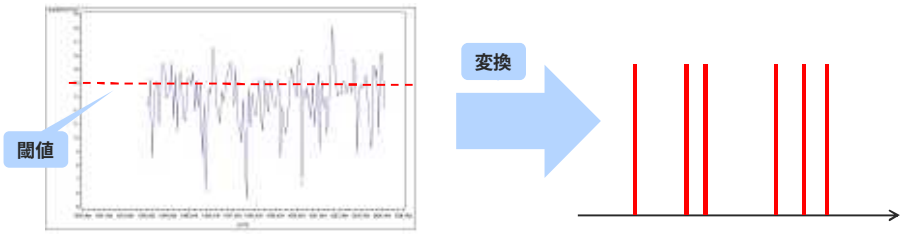
15

ステップ3：外れ値スコア時系列の構造の検出(2/4)

外れ値スコア時系列の構造の検出：

- ピークの検出：

1. ピークの検出：外れ値スコアが閾値を超えたら1、閾値を超えなければ0と変換。

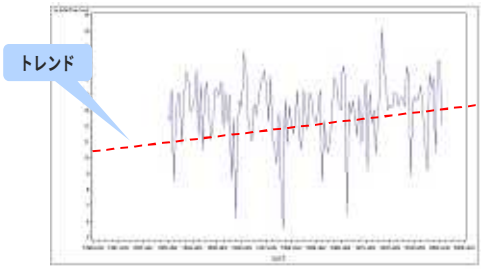


2. 周期の特定：1の間隔を統計的に分析し、ピークの出現周期を特定する。

ステップ3：外れ値スコア時系列の構造の検出(3/4)

外れ値スコア時系列の構造の検出：

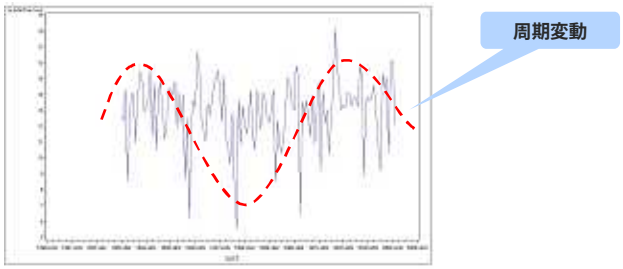
- トレンドの検出：線形回帰を用いる。



ステップ3：外れ値スコア時系列の構造の検出(4/4)

外れ値スコア時系列の構造の検出：

- 周期変動の検出：フーリエ変換を用いて特徴的な周波数を求める。



実験

実験データ

- AIRデータ：アメリカの飛行機の乗降客数の月次データ。

評価

- 以下の三つの場合について予測精度を比較した。
  - 外れ値を考慮しないARIMAモデル(FS)
  - 外れ値を考慮するARIMAモデル(FS)
  - 将来の外れ値（イベント）の位置を指定したARIMAモデル（提案手法）
- 予測精度の評価指標としてMAPEを用いた。

その他の設定

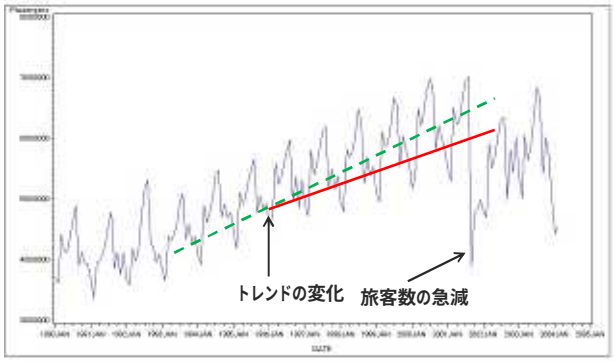
- 提案手法の予測とスコアリングにはForecast Serverに付属するSAS/High Performance Forecastingのproc hpfを使用した。

実験1: AIRデータからの非周期的ピークの検出(1/3)

入力時系列

AIRデータ：アメリカの飛行機の乗降客数の月次データ。

- 2001年9月に、911による大幅な旅客数減がみられる。
- 1996年6月に、目視では検出困難なトレンドの変化がある。

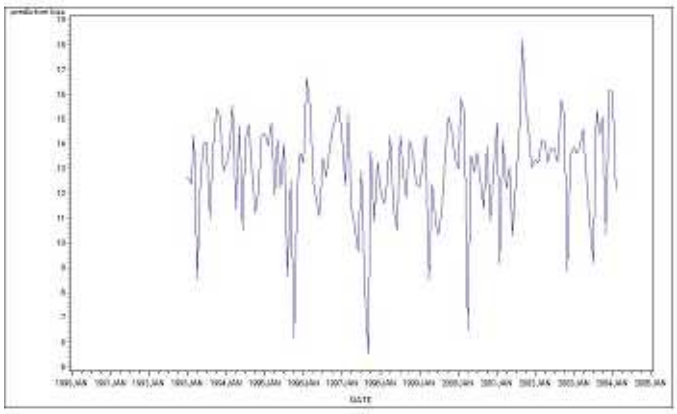


20

実験1: AIRデータからの非周期的ピークの検出(2/3)

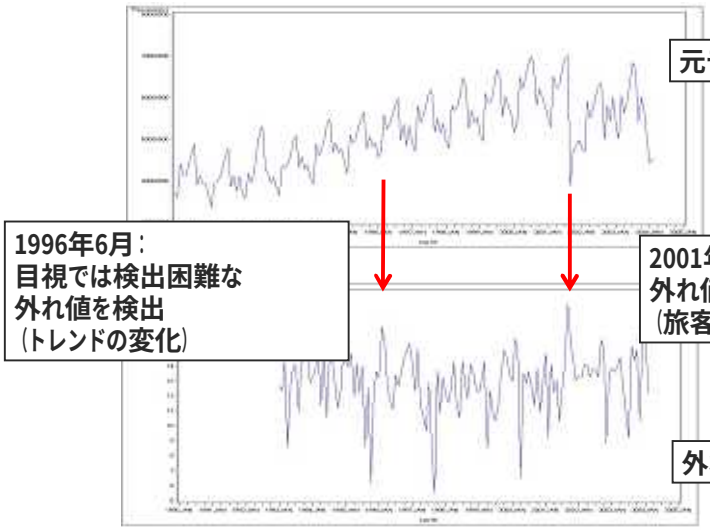
外れ値スコアの時系列

- 逐次的なモデル推定と予測によって、元の時系列を外れ値スコアの系列に変換



21

実験1: AIRデータからの非周期的ピークの検出(3/3)



非周期的な外れ値を自動的に検出できた

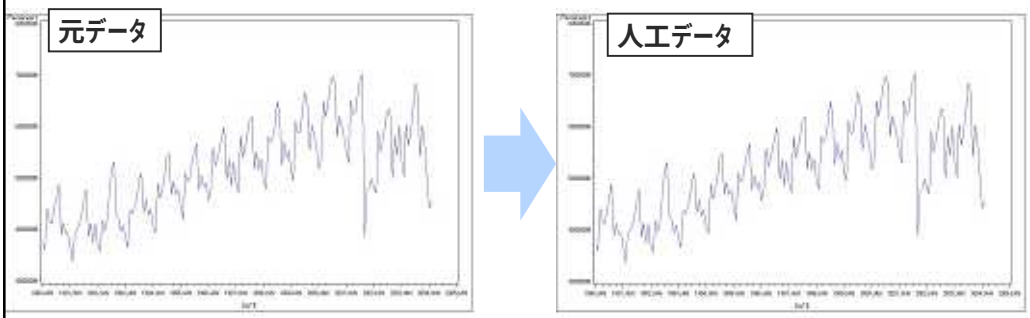
22

実験2: 人工データからの周期的ピークの検出(1/4)

入力時系列

AIRデータの値を、24ヶ月に一回30%大きくした人工データ。

- 24ヶ月周期で出現する外れ値があると期待される。

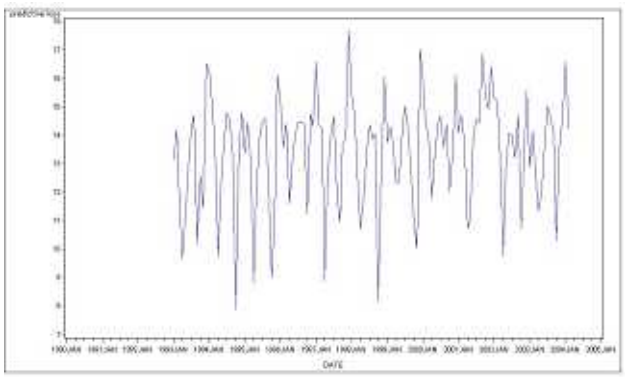


23

実験2: 人工データからの周期的ピークの検出(2/4)

外れ値スコアの時系列

- 逐次的なモデル推定と予測によって、元の時系列を外れ値スコアの系列に変換

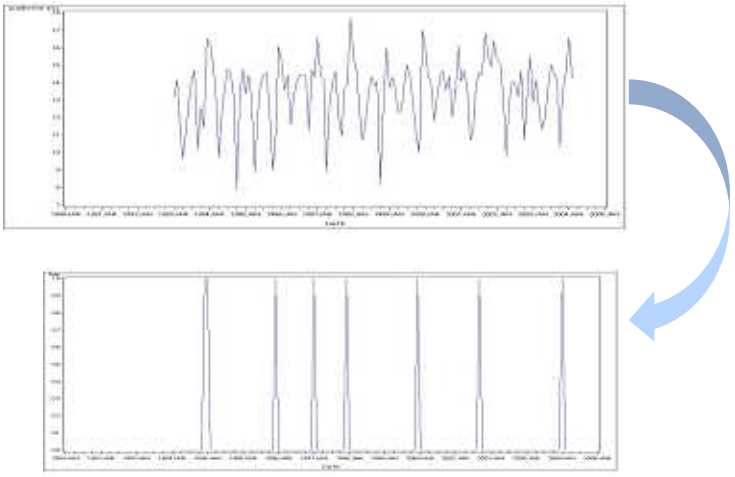


24

実験2: 人工データからの周期的ピークの検出(3/4)

周期の検出

- 外れ値スコア時系列を閾値との比較によって0,1の列に変換

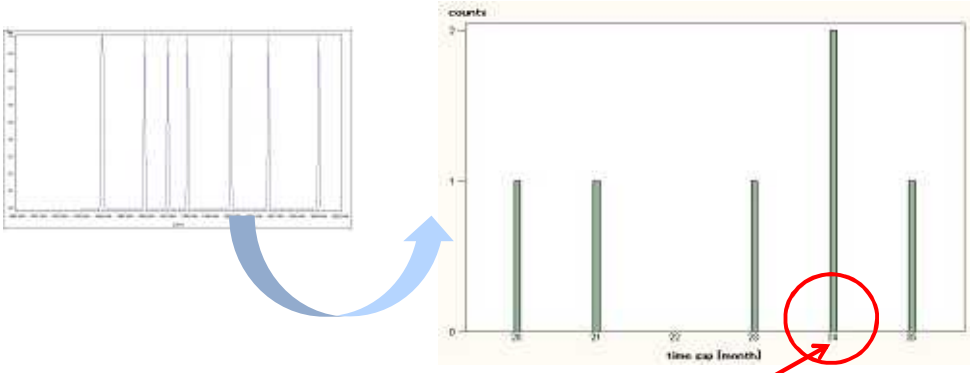


25

実験2: 人工データからの周期的ピークの検出(4/4)

周期の検出

- 0,1列から1の発生間隔を統計的な手法を用いて算出。



24ヶ月周期の周期性あり

周期的な外れ値を検出し、その周期を特定できた。

26

実験3: 人工データを用いた予測精度比較(1/3)

入力データ

- 前節と同じ人工データ。

評価

- 以下の三つの場合について予測精度を比較した。
  - 外れ値を考慮しないARIMAモデル(FS)
  - 外れ値を考慮するARIMAモデル(FS)
  - 将来の外れ値（イベント）の位置を指定したARIMAモデル（提案手法）
- 13年分の月次データの内、12年分を学習データとし、残りの1年分をテストデータとした。
- 予測精度の評価指標としてMAPEを用いた。

  MAPE: Mean Absolute Percent Errorの略。

  以下で定義され、予測精度が高いほどMAPEの値は小さくなる。

{ 評価期間中の各時点ごとの (| 実測値 – 予測値 | ÷ 実測値 × 100 (%)) の合計 } ÷ 評価期間

27

実験3: 人工データを用いた予測精度比較(2/3)

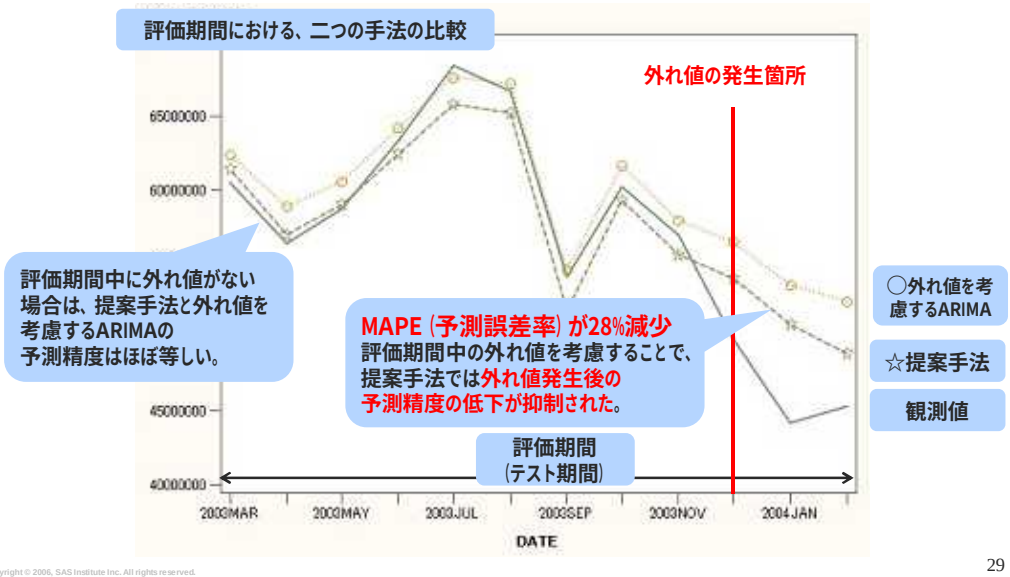
結果

モデル:	MAPE
外れ値を考慮しないARIMAモデル(FS)	10.37
外れ値を考慮するARIMAモデル(FS)	5.75
将来の外れ値（イベント）の位置を指定したARIMAモデル（提案手法）	4.15

MAPEが28%低下
（予測誤差が28%減少）

周期的な外れ値を自動的に検出して考慮することで、
予測精度がMAPEの比較で28%改善した。

実験3: 人工データを用いた予測精度比較(3/3)



まとめ

外れ値を含む時系列の予測問題

- 外れ値(時系列の通常の変動パターンから外れた例外的な値)を適切に取り扱うことで、時系列予測の精度を更に向上させたい。
- リスクの管理の観点から予測精度の向上は非常に重要。

SASの時系列予測プロシジャを用いた外れ値の構造の検出

- 「テストデータへの当てはめ統計量」(予測符号長)を用いてデータの外れ値度合いを外れ値スコアとして数値化。
- 外れ値スコア時系列の構造を分析し、将来の外れ値の出現位置(時刻)を考慮したモデルを構築。
- 将来に出現する外れ値を考慮することで従来手法と比較して予測精度が向上。

APPENDIX.
ARIMAモデルの概略

ARIMAモデル

・ARIMAモデルは時系列の時間発展モデルです。

・ARIMAモデルを用いると、時系列に含まれる以下の三つの成分をモデル化できます。

1. 平均値周りの振動（上がったら下がり、下がったら上がって平均の周りをうろうろする）成分
2. トレンド（平均値が上昇または下降する）成分
3. イベントの発生など、短期的にしか影響を及ぼさない突発的な成分

・ ARIMAモデルは、AR(Auto Regression)モデル、I(Integration)モデル、MA(Moving Average)モデルの三つのモデルを組み合わせたモデルです。以後のページでは、それぞれのモデルについて解説し、最後にそれを組み合わせたARIMAモデルについて説明します。

ARIMAモデル

ARモデル:AR(p)

従属変数（indonesia_demand; yと書く）の値が、直前p個の従属変数の値から決まるとするモデル：

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \ldots + a_p y(t-p) + \varepsilon(t)$$

・εはノイズを表します。

・ARモデルで推定されるパラメータはp個の係数a_1、a_2、…、a_p です。

・AR モデルは平均値周りでの振動（値が上がったら次は下がり、下がったら上がる、といった変動）を表します。

ARIMAモデル

Iモデル（階差モデル）：I(d)

長さdの期間における従属変数の変化量が、前の時刻の変化量から決まるというモデル：

$$y(t) - y(t-d) = c(y(t-1) - y(t-d-1)) + \varepsilon(t)$$

・εはノイズを表します。

・Iモデルで推定されるパラメータは1個の係数c です。

・I モデルは従属変数時系列のトレンド（平均値の上昇や下降）を表します。

ARIMAモデル

MAモデル：MA(q)

従属変数の値が現在と過去q個のノイズ（突発的なずれ）の大きさから決まるというモデル：

$$y(t) = \mu + b_0\varepsilon(t) + b_1\varepsilon(t-1) + \ldots + b_q\varepsilon(t-q)$$

・εはノイズを表します。

・MAモデルで推定されるパラメータはq個の係数b_0、b_1、…、b_q と平均項μです。

・MAモデルはイベントの発生の影響を受けて従属変数の値がきまるような状況を表しています。

ARIMAモデル：独立変数のMAモデル

独立変数を考慮する場合、ARIMAモデルでは以下のようなMAモデルを用いて従属変数の予測に独立変数を取り込みます。

独立変数のMAモデル:MA(q)

従属変数の値現在と過去q個の独立変数(coal_IMFなど；xと書く)変化量が、
前の時刻の変化量から決まるとする：

$$y(t) = \mu + b_0 x(t) + b_1 x(t-1) + \ldots + b_q x(t-q)$$

・独立変数を含むMAモデルで推定されるパラメータはq個の係数b_0、b_1、…、b_q と平均項μです。

・独立変数を含むMAモデルは過去の独立変数の値の影響を受けて現在の従属変数の値がきまるような状況を表しています。

ARIMAモデル

AR(Auto Regression)モデル、I(Integration)モデル、MA(Moving Average)モデルの三つのモデルを組み合わせるとARIMAモデルとなります。
ARIMAモデルの具体的な表式は以下のようになります。

$$y(t) - y(t-d) = \mu + \frac{\theta(\hat{B})}{\phi(\hat{B})}\varepsilon(t) + \frac{\omega(\hat{B})}{\delta(\hat{B})}x(t)$$

・yは従属変数を表します。

・xは独立変数を表します。

・dはIモデルのdと同様です。

・μはMAモデルの平均項と同様です。

・Bは（過去方向への）時間発展演算子で、以下のように定義されます。

$$\hat{B}y(t) = y(t-1)$$

ARIMAモデル

・θは移動平均演算子と呼ばれ、以下のような多項式で定義されます。

$$\theta(\hat{B}) = 1 - a_1\hat{B} - \ldots - a_p\hat{B}^p,$$

$$\rightarrow \theta(\hat{B})y(t) = y(t) - a_1 y(t-1) - \ldots - a_p y(t-p).$$

・Φは自己回帰演算子と呼ばれ、以下のような多項式で定義されます。

$$\phi(\hat{B}) = 1 - b_1\hat{B} - \ldots - b_q\hat{B}^q,$$

$$\rightarrow \phi(\hat{B})y(t) = y(t) - b_1 y(t-1) - \ldots b_q y(t-q).$$

ARIMAモデル

・ω もθと同様に移動平均演算子で、Bの多項式として定義されます。ただし、θと等しくはありません。

・δもΦと同様に自己回帰演算子で、Bの多項式として定義されます。ただし、Φとは等しくありません。

・独立変数に掛っている以下の分数を伝達関数と呼びます

$$\frac{\omega(\hat{B})}{\delta(\hat{B})}$$