4.7　表記と統計的検証に用いる代表的な指標

数値予報解説資料集で用いた表記と統計的検証に用いる代表的な指標などについて以下に説明する。

4.7.1　数値予報解説資料集で用いた表記

(1)　時刻の表記について

本資料集では、時刻を表記する際に、通常国内で用いられている日本標準時 (JST: Japan Standard Time) のほかに、協定世界時 (UTC: Coordinated Universal Time) を用いている。数値予報では国際的な観測データの交換やプロダクトの利用等の利便を考慮して、時刻は UTC で表記されることが多い。JST は UTC に対して 9 時間進んでいる。また、単に「時」を用いる場合は、日本標準時を意味する。

(2)　分解能の表記について

本資料集では、全球モデルの分解能について、xx を水平方向の切断波数、yy を鉛直層数として、“TxxLyy”[1] と表記することがある。また、セミラグランジアンモデルで線形格子 (北川 2005) を用いる場合は “TLxxLyy”[2] と、二次格子 (氏家ほか 2019) を用いる場合には “TQxxLyy”[3] と表記する。北緯 30 度において、TL959 は約 20 km 格子、TL479 は約 40 km 格子、TL319 は約 55 km 格子、TL159 は約 110 km 格子、TQ479 は約 27 km 格子、TQ319 は約 40 km 格子に相当する。

(3)　予測時間の表記について

数値予報では、統計的な検証や事例検証の結果を示す際に、予報対象時刻のほかに、初期時刻からの経過時間を予報時間 (FT: Forecast Time[4]) として表記している。

本資料集では、予報時間を

「予報時間」=「予報対象時刻」−「初期時刻」

で定義し、例えば、6 時間予報の場合、FT=6 と表記しており、時間の単位 [h] を省略している。

(4)　アンサンブル予報の表記について

アンサンブル予報では、複数の予測の集合（アンサンブル）を統計的に処理し、確率予測等の資料を作成する。本資料集では、予測の集合の平均を「アンサンブル平均」、個々の予測を「メンバー」と呼ぶ。また、摂動を加えているメンバーを「摂動ラン」、摂動を加えていないメンバーを「コントロールラン」と呼ぶ。全メンバーの数に対する、予測がある閾値を超える（または下回る）メンバーの数の割合を超過確率と呼ぶ。

(5)　緯度、経度の表記について

本資料集では、緯度、経度について、アルファベットを用いて例えば「北緯 40 度、東経 130 度」を「40°N, 130°E」、「南緯 40 度、西経 130 度」を「40°S, 130°W」などと略記する。

[1] T は三角形 (Triangular) 波数切断、L は層 (Level) を意味する。
[2] TL の L は線形 (Linear) 格子を意味する。
[3] TQ の Q は二次 (Quadratic) 格子を意味する。
[4] 英語圏では Forecast Range などと記述されることも多い。

4.7.2　統計的検証に用いる代表的な指標

(1)　平均誤差、二乗平均平方根誤差、誤差の標準偏差、改善率

予測誤差を表す基本的な指標として、平均誤差（ME: Mean Error、バイアスと表記する場合もある）と二乗平均平方根誤差 (RMSE: Root Mean Square Error) がある。これらは次式で定義される。

$$\mathrm{ME} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i) \tag{4.7.1}$$

$$\mathrm{RMSE} \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - a_i)^2} \tag{4.7.2}$$

ここで、N は標本数、x_i は予測値、a_i は実況値である。ME は予測値の実況値からの偏りの平均であり、0 に近いほど実況からのずれが小さいことを示す。RMSE は最小値の 0 に近いほど予測が実況に近いことを示す。

RMSE は ME の寄与とそれ以外を分離して、

$$\mathrm{RMSE}^2 = \mathrm{ME}^2 + \sigma_e^2 \tag{4.7.3}$$

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i - \mathrm{ME})^2 \tag{4.7.4}$$

と表すことができる。σ_e は誤差の標準偏差である。

本資料集では、予測に改良を加えた際の評価指標として、RMSE の改善率 (%) を用いる場合がある。RMSE の改善率は次式で定義される。

$$\mathrm{RMSE}\,改善率 \equiv \frac{\mathrm{RMSE_{cntl}} - \mathrm{RMSE_{test}}}{\mathrm{RMSE_{cntl}}} \times 100 \tag{4.7.5}$$

（RMSE 改善率 ≤ 100）

ここで、$\mathrm{RMSE_{cntl}}$ は基準となる予測の、$\mathrm{RMSE_{test}}$ は改良を加えた予測の RMSE である。

(2)　スプレッド

スプレッドは、アンサンブル予報のメンバーの広がりを示す指標であり、次式で定義される。

$$\text{スプレッド} \equiv \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}(x_{mn} - \overline{x_n})^2\right)} \tag{4.7.6}$$

ここで、M はアンサンブル予報のメンバー数、N は標本数、x_{mn} は m 番目のメンバーの予測値、$\overline{x_n}$ は

$$\overline{x_n} \equiv \frac{1}{M}\sum_{m=1}^{M} x_{mn} \tag{4.7.7}$$

で定義されるアンサンブル平均である。

(3) アノマリー相関係数

アノマリー相関係数 (ACC: Anomaly Correlation Coefficient) とは、予測値の基準値からの偏差（アノマリー）と実況値の基準値からの偏差との相関係数であり、次式で定義される。

$$\mathrm{ACC} \equiv \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(A_i - \overline{A}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{N}\left(A_i - \overline{A}\right)^2}} \quad (-1 \leq \mathrm{ACC} \leq 1) \qquad (4.7.8)$$

ただし、

$$X_i = x_i - c_i, \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (4.7.9)$$

$$A_i = a_i - c_i, \qquad \overline{A} = \frac{1}{N}\sum_{i=1}^{N} A_i \qquad (4.7.10)$$

である。ここで、N は標本数、x_i は予測値、a_i は実況値、c_i は基準値である。基準値としては気候値を用いる場合が多い。アノマリー相関係数は予測と実況の基準値からの偏差の相関を示し、基準値からの偏差の増減のパターンが完全に一致している場合には最大値の 1 をとり、相関が全くない場合には 0 をとり、逆に完全にパターンが反転している場合には最小値の −1 をとる。なお、アノマリー相関係数や ME, RMSE の解説は、梅津ほか (2013) に詳しい。

4.7.3 カテゴリー検証で用いる指標

カテゴリー検証では、まず、対象となる現象の有無を予測と実況それぞれについて判定し、その結果により標本を分類する。そして、それぞれのカテゴリーに分類された事例数を基に、予測の特性を検証するという手順を踏む。

(1) 分割表

分割表は、カテゴリー検証においてそれぞれのカテゴリーに分類された事例数を示す表（表 4.7.1）である。(2) から (12) に示す各スコアは、表 4.7.1 に示される各区分の事例数を用いて定義される。また、以下では全事例数を N=FO+FX+XO+XX、実況「現象あり」の事例数を M=FO+XO、実況「現象なし」の事例数を X=FX+XX と表す。

表 4.7.1 カテゴリー検証で用いる分割表。FO, FX, XO, XX はそれぞれの事例数を示す。

		実況		計
		あり	なし	
予測	あり	適中 (FO)	空振り (FX)	FO+FX
	なし	見逃し (XO)	適中 (XX)	XO+XX
計		M	X	N

(2) 適中率

適中率は、予測が適中した割合であり、次式で定義される。

$$\text{適中率} \equiv \frac{\mathrm{FO}+\mathrm{XX}}{N} \quad (0 \leq \text{適中率} \leq 1) \qquad (4.7.11)$$

最大値の 1 に近いほど予測の精度が高いことを示す。

(3) 空振り率

空振り率は、予測「現象あり」の事例数に対する空振り（予測「現象あり」かつ実況「現象なし」）の割合であり、次式で定義される。

$$\text{空振り率} \equiv \frac{\mathrm{FX}}{\mathrm{FO}+\mathrm{FX}} \quad (0 \leq \text{空振り率} \leq 1) \qquad (4.7.12)$$

最小値の 0 に近いほど空振り率が小さいことを示す。本資料集では分母を FO+FX としているが、代わりに N として定義する場合もある。

(4) 見逃し率

見逃し率は、実況「現象あり」の事例数に対する見逃し（実況「現象あり」かつ予測「現象なし」）の割合であり、次式で定義される。

$$\text{見逃し率} \equiv \frac{\mathrm{XO}}{M} \quad (0 \leq \text{見逃し率} \leq 1) \qquad (4.7.13)$$

最小値の 0 に近いほど見逃し率が小さいことを示す。本資料集では分母を M としているが、代わりに N として定義する場合もある。

(5) 捕捉率

捕捉率 (H_r: Hit Rate、POD(Probability Of Detection) とも呼ばれる) は、実況「現象あり」のときに予測が適中した割合であり、次式で定義される。

$$H_r \equiv \frac{\mathrm{FO}}{M} \quad (0 \leq H_r \leq 1) \qquad (4.7.14)$$

最大値の 1 に近いほど見逃し率が小さいことを示す。捕捉率は、ROC 曲線 4.7.4 (5) のプロットに用いられる。

(6) 体積率

体積率 (V_r: Volume Ratio) は、全事例のうち予測の「現象あり」の事例の割合を示す。

$$V_r \equiv \frac{\mathrm{FO}+\mathrm{FX}}{N} \qquad (4.7.15)$$

複数の予測の捕捉率が等しい場合、体積率が小さい予測ほど空振り率が小さい良い予測と言える。

(7)　誤検出率

誤検出率 (F_r: False Alarm Rate) は、実況「現象なし」のときに予測が外れた割合である。空振り率 (4.7.12) 式とは分母が異なり、次式で定義される。

$$F_r \equiv \frac{\mathrm{FX}}{X} \quad (0 \leq F_r \leq 1) \tag{4.7.16}$$

最小値の 0 に近いほど、誤検出率が小さく予測の精度が高いことを示す。誤検出率は捕捉率 (5) とともに ROC 曲線 4.7.4 (5) のプロットに用いられる。

(8)　バイアススコア

バイアススコア (BI: Bias Score) は、実況「現象あり」の事例数に対する予測「現象あり」の事例数の比であり、次式で定義される。

$$\mathrm{BI} \equiv \frac{\mathrm{FO}+\mathrm{FX}}{M} \quad (0 \leq \mathrm{BI}) \tag{4.7.17}$$

予測と実況で「現象あり」の事例数が一致する場合に 1 となる。1 より大きいほど予測の「現象あり」の頻度が過大、1 より小さいほど予測の「現象あり」の頻度が過小であることを示す。

(9)　気候学的出現率

現象の気候学的出現率 P_c は、標本から見積もられる「現象あり」の平均的な出現確率であり、次式で定義される。

$$P_c \equiv \frac{M}{N} \quad (0 \leq P_c \leq 1) \tag{4.7.18}$$

この量は実況のみから決まり、予測の精度にはよらない。予測の精度を評価する際の基準値の設定にしばしば用いられる。

(10)　スレットスコア

スレットスコア (TS: Threat Score) は、予測または実況で「現象あり」の場合の予測適中事例数に着目して予測精度を評価する指標であり、次式で定義される。

$$\mathrm{TS} \equiv \frac{\mathrm{FO}}{\mathrm{FO}+\mathrm{FX}+\mathrm{XO}} \quad (0 \leq \mathrm{TS} \leq 1) \tag{4.7.19}$$

出現頻度の低い現象（$N \gg M$、したがって、XX≫FO, FX, XO となって、予測「現象なし」による寄与だけで適中率が 1 に近い現象）について XX の影響を除いて検証するのに有効である。本スコアは最大値の 1 に近いほど予測の精度が高いことを示す。なお、スレットスコアは現象の気候学的出現率の影響を受けやすく、異なる標本や出現率の異なる現象に対する予測の精度を比較するのには適さない。この問題を緩和するため、次項で説明するエクイタブルスレットスコアなどが考案されている。

(11)　エクイタブルスレットスコア

エクイタブルスレットスコア (ETS: Equitable Threat Score) は、前項のスレットスコアが現象の気候学的出現率の影響を受けやすいため、気候学的な確率で「現象あり」が適中した頻度を除いて求めたスレットスコアであり、次式で定義される (Schaefer 1990)。

$$\mathrm{ETS} \equiv \frac{\mathrm{FO}-S_f}{\mathrm{FO}+\mathrm{FX}+\mathrm{XO}-S_f} \quad \left(-\frac{1}{3} \leq \mathrm{ETS} \leq 1\right) \tag{4.7.20}$$

ただし、

$$S_f = P_c(\mathrm{FO}+\mathrm{FX}) \tag{4.7.21}$$

である。ここで、S_f は「現象あり」をランダムに FO+FX 回予測した場合（ランダム予測）の「現象あり」の適中事例数である。本スコアは、最大値の 1 に近いほど予測の精度が高いことを示す。また、ランダム予測で 0 となり、FO=XX=0, FX=XO=$N/2$ の場合に最小値 −1/3 をとる。

(12)　スキルスコア

スキルスコア (Skill Score) は気候学的確率などによる予測の難易を取り除いて、予測の技術力を評価する指数であり、一般に次式のように定義される。

$$\text{スキルスコア} \equiv \frac{S_{\mathrm{fcst}}-S_{\mathrm{ref}}}{S_{\mathrm{pfct}}-S_{\mathrm{ref}}} \tag{4.7.22}$$

ここで、$S_{\mathrm{fcst}}, S_{\mathrm{pfct}}, S_{\mathrm{ref}}$ は、評価対象の予測・完全予測・比較の基準となる予測（気候学的確率など）の各スコア（適中率）である。本スコアは、最大値の 1 に近いほど予測の精度が高いことを示し、比較の基準となる予測よりも精度が劣る場合、負の値となる。

代表的なスキルスコアは Heidke のスキルスコア (HSS: Heidke Skill Score) で、気候学的な確率で「現象あり」および「現象なし」が適中した頻度を除いて求める適中率であり、次式で定義される。

$$\mathrm{HSS} \equiv \frac{\mathrm{FO}+\mathrm{XX}-S}{N-S} \quad (-1 \leq \mathrm{HSS} \leq 1) \tag{4.7.23}$$

ただし、

$$S = P_c(\mathrm{FO}+\mathrm{FX}) + P_x(\mathrm{XO}+\mathrm{XX}), \quad P_x = \frac{X}{N} \tag{4.7.24}$$

である。ここで、P_x は「現象なし」の気候学的出現率、S は「現象あり」を FO+FX 回（すなわち、「現象なし」を残りの XO+XX 回）ランダムに予測した場合（ランダム予測）の適中事例数である。HSS は、最大値の 1 に近づくほど精度が高く、ランダム予測で 0 となり、FO=XX=0, FX=XO=N/2 の場合に最小値 −1 をとる。前項のエクイタブルスレットスコアもスキルスコアの一つで、Gilbert Skill Score とも呼ばれている。

(13) Roebber ダイアグラム

Roebber (2009) はカテゴリ検証による複数のスコア（捕捉率、空振り率、バイアススコア、スレットスコア）を一つのグラフに表す方法を考案した。検証結果を縦軸に捕捉率 (POD: Probability Of Detection)、横軸に 1− 空振り率 (SR: Success Ratio) をとってプロットすると、捕捉率と空振り率から BI と TS が計算できるため、等値線を目安にバイアススコアとスレットスコアも確認できるグラフとなる（図 4.7.1）。本資料集では、これを Roebber ダイアグラムと呼ぶ。各スコアが 1 に近づくほど（グラフの右上へ近づくほど）、良い予測となる。このグラフでは 4 つのスコアを一目で確認でき、予測特性の変化を把握しやすい。特に、バイアススコアとスレットスコアの変化を捕捉率と空振り率の変化で説明することが容易となる。

例えば、図 4.7.1 の①のようにスコアが変化する場合、捕捉率、空振り率、バイアススコア、スレットスコアのいずれも改善となる。これに対し②の場合には、①と同様にバイアススコア、スレットスコアとも改善しているが、空振り率が増加している。空振り率が大きいにもかかわらず、バイアススコア・スレットスコアが改善している理由は、捕捉率の増加の割合が空振り率の増加に比べて大きいためである。このように①と②ではいずれもバイアススコアとスレットスコアがともに改善しているが、本グラフを用いることで予測の変化傾向の違い（捕捉率と空振り率の変化の違い）が一目で確認できる。

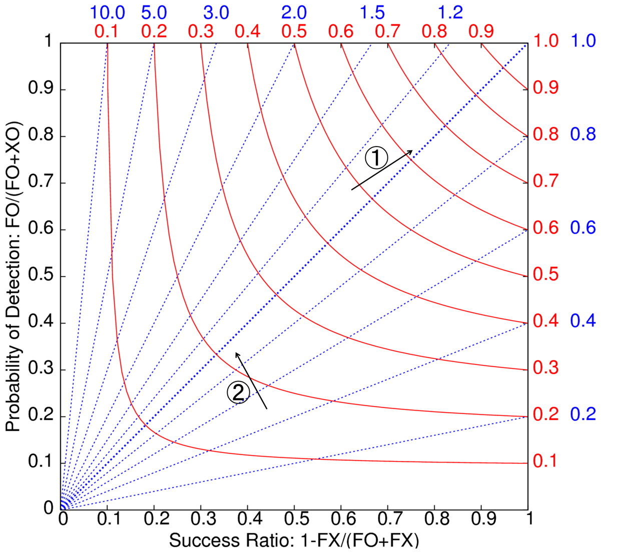

図 4.7.1 Roebber ダイアグラムの模式図。横軸は 1-空振り率、縦軸は捕捉率、青の破線はバイアススコアの、赤の実線はスレットスコアの各等値線。

(14) FSS

FSS(Fractions Skill Score) は、現象の表現に空間的な曖昧さを与えて評価する検証スコアである（Roberts and Lean 2008 参照、幾田 2010 に詳しい）。

平面上のある変量の観測の分布を O_r、予報の分布を F_r とする。変量は任意の閾値 q で 2 値化でき、2 値化した観測を I_O、予報を I_F とすると、次式のように表せる。

$$I_O = \begin{cases} 1 & O_r \geq q \\ 0 & O_r < q \end{cases} \tag{4.7.25}$$

$$I_F = \begin{cases} 1 & F_r \geq q \\ 0 & F_r < q \end{cases} \tag{4.7.26}$$

この 2 値化した変量を用いた検証は空間的な位置ずれを許容せず、検証格子のスケールでの適合を厳密に検証することを意味する。

次に、この I_O と I_F に空間スケールを考慮し、分布の適合の判定に曖昧さを追加するため、分数化を行う。具体的には、検証対象格子を中心とする 1 辺 n 格子の正方形領域を考え、この正方形領域に含まれる 2 値化した格子情報を次式に従って領域平均する。

$$\begin{aligned} O(n)_{i,j} &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} I_O\left[i+k-1-\frac{n-1}{2},\right. \\ &\qquad \left. j+l-1-\frac{n-1}{2}\right]\cdot K(n)_{k,l} \\ F(n)_{i,j} &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} I_F\left[i+k-1-\frac{n-1}{2},\right. \\ &\qquad \left. j+l-1-\frac{n-1}{2}\right]\cdot K(n)_{k,l} \end{aligned} \tag{4.7.27}$$

ここで $O(n)$ と $F(n)$ は分数化した観測と予報、添字の i, j は格子番号である。また、$K(n)$ はカーネル関数で一般的にはガウシアンカーネルなどが考えられるが、ここでは格子内平均を取り扱うためカーネル関数は一様とする。

分数化した変量 $O(n)$ と $F(n)$ によって二乗平均誤差 (MSE) が次式によって計算される。

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[O(n)_{i,j} - F(n)_{i,j}]^2 \tag{4.7.28}$$

ここで、N_x と N_y は検証領域の x 方向の格子数と y 方向の格子数である。ここでは、簡単のため検証領域は矩形領域であると仮定している。

FSS は分数化された観測 $O(n)$ と予報 $F(n)$ によって記述される MSE のスキルスコアであるため、予報スキルを評価するための相対的な基準となる参照値が必要である。FSS の参照値は、$O(n)$ と $F(n)$ を用いて次

式のように定義される。

$$\mathrm{MSE}_{(n)\mathrm{ref}} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O^2(n)_{i,j} + F^2(n)_{i,j}] \tag{4.7.29}$$

この参照値 $\mathrm{MSE}_{(n)\mathrm{ref}}$ は、任意の MSE の取りうる最大の値であり、予報と観測の総数が検証領域の格子数を超えない場合において、予報と観測の適合が無い場合の MSE に相当する。

FSS は、分数化した観測と予報によって記述される $\mathrm{MSE}_{(n)}$、その参照値である $\mathrm{MSE}_{(n)\mathrm{ref}}$、そして完全予報の $\mathrm{MSE}_{(n)\mathrm{perfect}}(=0)$ を用いて次式で定義される。

$$\mathrm{FSS}_{(n)} = \frac{\mathrm{MSE}_{(n)} - \mathrm{MSE}_{(n)\mathrm{ref}}}{\mathrm{MSE}_{(n)\mathrm{perfect}} - \mathrm{MSE}_{(n)\mathrm{ref}}} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n)\mathrm{ref}}} \tag{4.7.30}$$

この式から分かるように FSS は 0 から 1 の値をとり、1 で完全予報、0 で観測と予報の適合がまったく無い場合となる。

4.7.4　確率予測に関する指標など

(1)　ブライアスコア

ブライアスコア (BS: Brier Score) は、確率予測の統計検証の基本的指標である。ある現象の出現確率を対象とする予測について、次式で定義される。

$$\mathrm{BS} \equiv \frac{1}{N} \sum_{i=1}^{N} (p_i - a_i)^2 \quad (0 \leq \mathrm{BS} \leq 1) \tag{4.7.31}$$

ここで、p_i は確率予測値（0 から 1）、a_i は実況値（現象ありで 1、なしで 0）、N は標本数である。BS は完全に適中する決定論的な（p_i=0 または 1 の）予測（完全予測と呼ばれる）で最小値の 0 をとり、0 に近いほど予測の精度が高いことを示す。また、現象の気候学的出現率 P_c(4.7.18) 式を常に確率予測値とする予測（気候値予測と呼ばれる）のブライアスコア $\mathrm{BS_c}$ は

$$\mathrm{BS_c} \equiv P_c(1 - P_c) \tag{4.7.32}$$

となる。ブライアスコアは、現象の気候学的出現率の影響を受けるため、異なる標本や出現率の異なる現象に対する予測の精度を比較するのには適さない。例えば上の $\mathrm{BS_c}$ は P_c 依存性を持ち、同じ予測手法（ここでは気候値予測）に対しても P_c の値に応じて異なる値をとる (Stanski et al. 1989)。この問題を緩和するため、次項で説明するブライアスキルスコアが考案されている。

(2)　ブライアスキルスコア

ブライアスキルスコア (BSS: Brier Skill Score) は、ブライアスコアに基づくスキルスコアであり、通常気候値予測を基準とした予測の改善の度合いを示す。本スコアは、ブライアスコア BS、気候値予測によるブライアスコア $\mathrm{BS_c}$ を用いて

$$\mathrm{BSS} \equiv \frac{\mathrm{BS_c} - \mathrm{BS}}{\mathrm{BS_c}} \quad (\mathrm{BSS} \leq 1) \tag{4.7.33}$$

で定義され、完全予測で 1、気候値予測で 0、気候値予測より誤差が大きいと負となる。

(3)　Murphy の分解

Murphy (1973) は、ブライアスコアと予測の特性との関連を理解しやすくするため、ブライアスコアを信頼度 (Reliability)、分離度 (Resolution)、不確実性 (Uncertainty) の 3 つの項に分解した。これを Murphy の分解と呼ぶ（高野 2002 などに詳しい）。

確率予測において、確率予測値を L 個の区間に分け、標本を確率予測値の属する区間に応じて分類することを考える。確率予測値が l 番目の区間に属する標本数を N_l ($N = \sum_{l=1}^{L} N_l$)、このうち実況が「現象あり」であった事例数を M_l ($M = \sum_{l=1}^{L} M_l$)、確率予測値の l 番目の区間の区間代表値を p_l とすると、Murphy の分解によりブライアスコアは以下のように表される。

$$\mathrm{BS} = \text{信頼度} - \text{分離度} + \text{不確実性} \tag{4.7.34}$$

$$\text{信頼度} = \sum_{l=1}^{L} \left(p_l - \frac{M_l}{N_l} \right)^2 \frac{N_l}{N} \tag{4.7.35}$$

$$\text{分離度} = \sum_{l=1}^{L} \left(\frac{M}{N} - \frac{M_l}{N_l} \right)^2 \frac{N_l}{N} \tag{4.7.36}$$

$$\text{不確実性} = \frac{M}{N} \left(1 - \frac{M}{N} \right) \tag{4.7.37}$$

信頼度は、確率予測値 (p_l) と実況での現象の出現相対頻度 (M_l/N_l) が一致すれば最小値の 0 となる。分離度は、確率予測値に対応する実況での現象の出現相対頻度 (M_l/N_l) が気候学的出現率 ($P_c = M/N$) から離れているほど大きい値をとる。不確実性は、現象の気候学的出現率のみによって決まり、予測の手法にはよらない。例えば、$P_c = 0.5$ の場合に不確実性は最大値の 0.25 をとる。また、不確実性=$\mathrm{BS_c}$ が成り立つ。これらを用いて、ブライアスキルスコアを次のように書くことができる。

$$\mathrm{BSS} = \frac{\text{分離度} - \text{信頼度}}{\text{不確実性}} \tag{4.7.38}$$

(4)　確率値別出現率図

確率値別出現率図（Reliability Diagram, Attributes Diagram とも呼ばれる）は、予測された現象出現確率 P_{fcst} を横軸に、実況で現象が出現した相対頻度 P_{obs} を縦軸にとり、確率予測の特性を示した図である（図 4.7.2 参照、Wilks 2011 などに詳しい）。一般に、確率予測の特性は確率値別出現率図上で曲線として表される。この曲線を信頼度曲線 (Reliability curve) と呼ぶ。

信頼度曲線の特性は、Murphy の分解 (3) の信頼度、分離度と関連付けることができる。横軸 P_{fcst} の各値について、信頼度（あるいは分離度）への寄与は、信頼度曲線上の点から対角線 $P_{\mathrm{obs}}=P_{\mathrm{fcst}}$（理想直線）上の点（あるいは直線 $P_{\mathrm{fcst}}=P_c$ 上の点）までの距離の二乗として表現される。P_{fcst} の各値でのこれらの寄与を、標本数に比例する重みで平均して信頼度（あるいは分離度）が得られる。例えば、no-skill line（直線 $P_{\mathrm{obs}} = (P_{\mathrm{fcst}} + P_c)/2$）上の点では、信頼度と分離度への寄与は等しい大きさを持ち、ブライアスキルスコアへの寄与が 0 となる。また no-skill line と直線 $P_{\mathrm{fcst}} = P_c$ との間の領域（分離度への寄与 > 信頼度への寄与、図 4.7.2 灰色の領域）内に位置する点は、ブライアスキルスコアに正の寄与を持つ。

特別な場合として、気候値予測 4.7.4 (1) では 1 点 $(P_{\mathrm{fcst}}, P_{\mathrm{obs}}) = (P_c, P_c)$ が信頼度曲線に対応する。また、次の 2 つの特性を示す確率予測は精度が高い。

- 信頼度曲線が対角線に（信頼度への寄与が最小値の 0 に）近い。
- 信頼度曲線上の大きい標本数に対応する点が点 $(P_{\mathrm{fcst}}, P_{\mathrm{obs}}) = (P_c, P_c)$（気候値予測）から離れた位置（確率値別出現率図の左下または右上寄り）に分布する（分離度が大きい）。

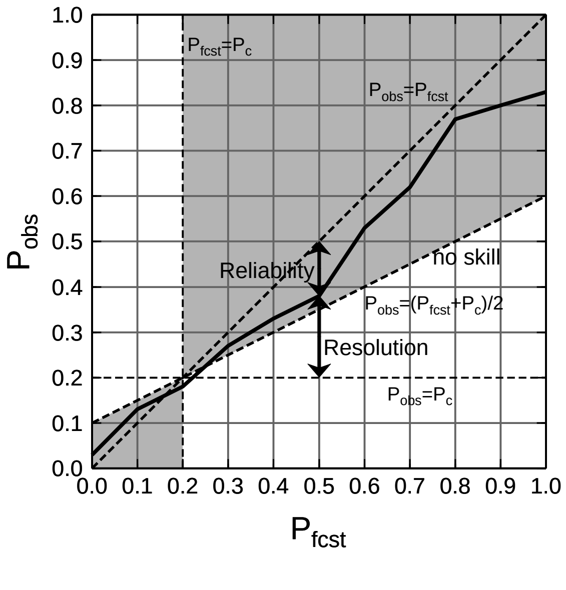

図 4.7.2　確率値別出現率図の模式図。横軸は予測現象出現確率、縦軸は実況現象出現相対頻度、実線が信頼度曲線である。対角線、直線 $P_{\mathrm{obs}} = P_c$ との差の二乗がそれぞれ信頼度 (Reliability)、分離度 (Resolution) への寄与に対応している。灰色の領域内の点はブライアスキルスコアに正の寄与を持つ。

(5)　ROC 曲線、ROC 面積、ROC 面積スキルスコア

現象の予測出現確率にある閾値を設定し、これを予測の「現象あり」「現象なし」を判定する基準とすることが可能である。様々な閾値それぞれについて作成した分割表を基に、閾値が変化したときの F_r–H_r 平面上の軌跡をプロットしたものが ROC 曲線（ROC curve: Relative Operating Characteristic curve、相対作用特性曲線）である（図 4.7.3 参照、高野 2002 などに詳しい）。平面内の左上方の領域では $H_r > F_r$ であり、平面の左上側に膨らんだ ROC 曲線特性を持つ確率予測ほど精度が高いものと見なせる。したがって、ROC 曲線から下の領域（図 4.7.3 灰色の領域）の面積（ROCA: ROC Area、ROC 面積）は、情報価値の高い確率予測ほど大きくなる。ROC 面積スキルスコア (ROCASS: ROC Area Skill Score) は、情報価値のない予測 ($H_r = F_r$) を基準として ROC 面積を評価するものであり、次式で定義される。

$$\mathrm{ROCASS} \equiv 2(\mathrm{ROCA} - 0.5) \quad (-1 \leq \mathrm{ROCASS} \leq 1) \tag{4.7.39}$$

本スコアは、完全予測で最大値の 1 をとる。また、情報価値のない予測（例えば、区間 [0, 1] から一様ランダムに抽出した値を確率予測値とする予測など）では 0 となる。

(6)　CRPS

CRPS (Continuous Ranked Probability Score) は、確率予測の統計検証の指標の 1 つである。連続物理量

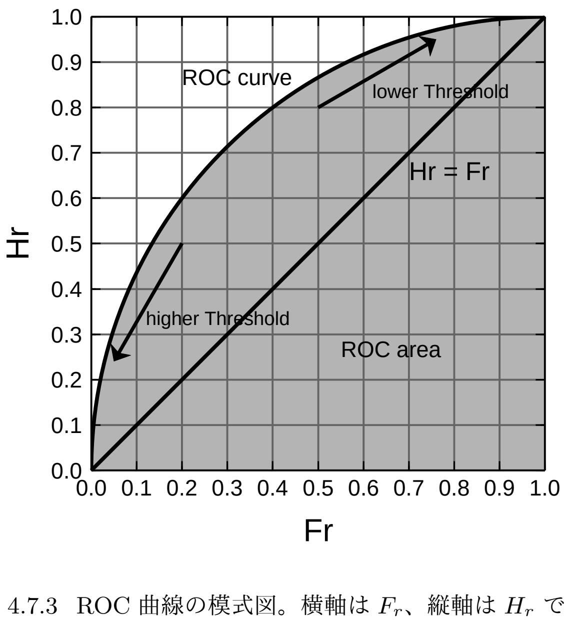

図 4.7.3　ROC 曲線の模式図。横軸は F_r、縦軸は H_r である。灰色の領域の面積が ROC 面積である。

x に対する CRPS は次式で定義される。

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left[P_i(x) - A_i(x)\right]^2 dx$$
$$(0 \leq \mathrm{CRPS}) \qquad (4.7.40)$$

ここで、N は標本数、P_i と A_i はそれぞれ予測と実況の累積分布関数であり、次式で定義される。

$$P_i(x) = \int_{-\infty}^{x}\rho_i(x')\,dx' \qquad (4.7.41)$$

$$A_i(x) = H(x - a_i) \qquad (4.7.42)$$

ここで、ρ_i は予測された確率密度関数、a_i は実況値、$H(x)$ は階段関数である。

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \qquad (4.7.43)$$

CRPS は完全に適中する決定論的な予測で最小値 0 をとり、0 に近いほど予測の精度が高いことを示す。単位は物理量 x と同じである。

また、物理量 x が閾値 t 以下となる現象の確率予測に対するブライアスコアを $\mathrm{BS}(t)$ とおくと、

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}\mathrm{BS}(t)dt \qquad (4.7.44)$$

の関係がある。

参考文献

幾田泰酵, 2010: 高分解能モデルの降水予報精度評価に適した検証手法. 平成 22 年度数値予報研修テキスト, 気象庁予報部, 11–17.

梅津浩典, 室井ちあし, 原旅人, 2013: 検証指標. 数値予報課報告・別冊第 59 号, 気象庁予報部, 6–15.

北川裕人, 2005: 全球・領域・台風モデル. 平成 17 年度数値予報研修テキスト, 気象庁予報部, 38–43.

高野清治, 2002: アンサンブル予報の利用技術. 気象研究ノート, **201**, 73–103.

氏家将志, 堀田大介, 黒木志洸, 2019: スペクトラルブロッキングの軽減. 数値予報課報告・別冊第 65 号, 気象庁予報部, 25–29.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

Roberts, N. M. and H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Mon. Wea. Rev.*, **136**, 78–97.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601–608.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. *Research Rep.*, **89-5**, Forecast Research Division, Atmospheric Envirnment Service, Environment Canada, 114 pp.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, International Geophysics, Vol. 100. Academic Press, 334-340 pp.