気象環境による体調予測システム

Abstract

気象環境や働き方の変化の中、西洋医学では解明できない体調不良や違和感を持つ人が増えている。一般論として気圧や気温など気象要素単体と体調との関係が示されているが、それに適合しない人も多く、その解決手段として東洋医学に解決を求める事例も増えている。そこで、東洋医学に代表される鍼治療に通う患者に対し、症状ごとのアンケート結果と、その期間の気象データから各気象要素の相互間の関係を統計的に解釈し、学習モデルを生成、気象数値予報モデルのデータを用いて体調変化を確率値として提供するシステムを作成した。これにより、自身の気象環境と体調変化の関係を意識し、また、その予防に繋げることができると確信する

初版：2023年6月11日

自己紹介

氏名：植田忠行

職務履歴：事務機器メーカで印刷業界向け電子写真プリンターのエレキ制御系開発（回路設計、駆動制御設計、規格対応など）を担当

所有スキル：気象予報士、第1種情報処理技術者、統計検討準1級、気象データアナリスト試験講座受講満了、駆動制御理論、基板回路設計など

参加組織：日本気象予報士会社員会員、気象ビジネス推進コンソーシアム（新規気象ビジネス創出WG会員）

目的

本体調予測システムは、気象病と言われている症状に対し、小沢鍼灸院の多くの患者様、また協力頂いた多くの方々の貴重な個人の体調データを基に、気象要素との関係を統計的に解析し、独自のアルゴリズムを通して10日までの体調予測を行うものである。

ここでは、体調予測システムの構成、また体調予測のモデル生成から予測値を算出するまでの過程を下記の項目に分けて説明する

1．予測モデル生成
　1-1.気象データ収集と処理（説明変数生成）
　1-2.体調レベル標本の処理（目的変数生成）

2．予測アルゴリズム生成
　2-1.予測アルゴリズムの構成
　2-2.予測モデルの生成

3．体調予測値の算出
　3-1.気象数値予測データの収集
　3-2.体調予測値の算出

4．考察と課題、謝辞

本システムの解析フローをFig. 1 に示す。

モデル生成のプロセス

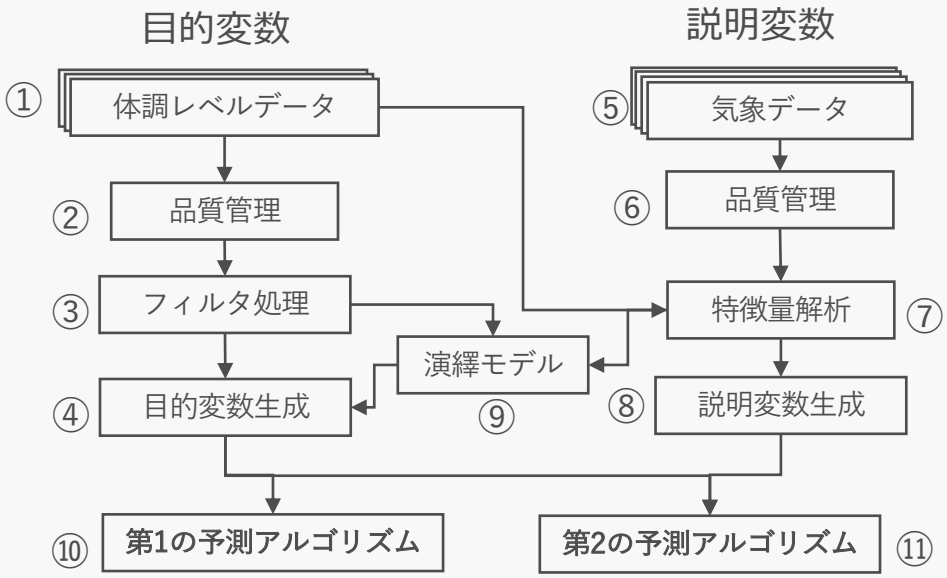


Fig. 1

番号	内容
①	1から4でカテゴリ化した体調レベル
②	欠損値や異常値を補正
③	体調レベルデータの平滑化
④	目的変数行列生成
⑤	気象庁のアメダスから得た気象データ
⑥	気象庁提供の信頼度評価
⑦	気象データの特徴量算出
⑧	説明変数行列生成
⑨	特徴量をとらえた演繹モデル
⑩	予測アルゴリズム１生成
⑪	予測アルゴリズム２生成

1．予測モデル生成

1-1.気象データ収集と処理（説明変数生成）

ここではモデル生成に必要な説明変数を生成する過程を説明する

体調レベルを説明する気象データは、日々気象庁から配信されるアメダス（AMeDAS：Automated Meteorological Data Acquisition System：自動気象データ収集システム）から収集を行う。

記号	内容	記号	内容
T_{max}	1 日の最高気温	RH_{ave}	1 日の平均相対湿度
T_{min}	1 日の最低気温	RH_{min}	1 日の最低相対湿度
P_{ave}	1 日の平均海面気圧	WP_{ave}	1 日の平均水蒸気圧
P_{min}	1 日の最低海面気圧	WS_{ave}	1 日の平均風速
WE_{ave}	1 日の代表天気	WB_{mod}	1 日の最頻風向

次に、配信されたデータに対し品質管理を行う。これは気象庁が同時に提供しているデータの信頼度（※1）を参考にして過去の系統ずれを補正し、確からしいパラメータとするためである

※1: 観測環境などの変化の前後で、値が不均質となったデータの扱い、欠損データなど

パラメータの種類によっては、正規性が確保できないものもあるため、シャピロ-ウィルクの正規性検定(1)などを用いて数値を評価し、必要であれば変数変換を行う。

$$\text{検定統計量 } W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \cdots\cdots (1)$$

頻繁に用いる手法としては、対数変換(2)がある、これは偏りを持った変数に対し、おおよそ正規分布に変換するものである。線形解析などは誤差に正規分布を前提にしていることが理由である。

$$x_{ti} = log x_{tj} \quad \cdots\cdots (2)$$

次に、補正された気象データに対し気象変化を表す特徴量を導く。

気象病による体調レベルは、時点の気象パラメータ値によるものより、数日前からの変化の累積値が影響していることが特徴量解析により分かった。ただし、年齢や性別による変化も大きく、一概には言えないため、ここでは平均値をとっている。

これによりアメダスの気象パラメータから別の特徴量を作成することで、体調レベルを精度良く予測することが出来きた。特徴量変換として代表的なものが重みづけ移動平均(3)である

$$\tilde{x}_t = \sum_{i=t-n}^{t} \omega_i x_i \Big| \quad \sum_{i=t-n}^{t} \omega_i = 0 \quad \cdots\cdots (3)$$

ただし特徴量については季節差などがあり、気象学的に有意があると考えられるパラメータに対し実行している。

以上により、アメダス配信データから、品質管理を経て、予測アルゴリズムモデルに入力する説明変数行列(4)を生成することができる

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} \quad \cdots\cdots (4)$$

1-2.体調レベル標本の処理（目的変数生成）

ここからはモデル生成に必要な目的変数を生成する過程を説明する

最初に、観測した体調レベル標本（1良い~4悪い）を表すカテゴリデータについて品質管理を行う。主な処理として、欠損値の判別、異常データの処理を行う

欠損値については、平均値代入、回帰代入、比率補完などがあるが、ここでは平均値代入を用いている。同じ日時に対して他の体調レベル値との平均を算出し、欠損箇所に代入するものである

次に、フィルター処理(5)を行う。フィルター処理を行う理由は、体調レベルの変化は、時点の状況のみで決まるものではなく連続的に変化していることにある。この傾向は特徴量解析から導いた。

$$\tilde{y}_t = \sum_{i=t-n}^{t} \omega_i y_i \,\Big|\, \sum_{i=t-n}^{t} \omega_i = 0 \qquad \cdots\cdots (5)$$

さて、誤差の大きい体調レベルに対し、確からしい目的変数を生成する必要がある
そこで、観測した体調レベルと、気象パラメータの関係を傾向的に解析し、最終的な目的変数を生成する。

ここで気象予報士の知見が必要となる。まずは多くの気象データから現象の変化を示す気象要素を気象メカニズム的に捉える（※1）。体調レベルはその生成メカニズムが不明なため、BlackBoxとしてデータドリブンで気象要素の特徴量を抽出する。利用した手法は、木構造のアルゴリズムを用いてモデル化し、体調レベルに対する寄与度はSHAP法（Fig. 2 ）を使い、そこから抽出された特徴量から、体調レベルを導出する演繹モデル（※2）を作成した

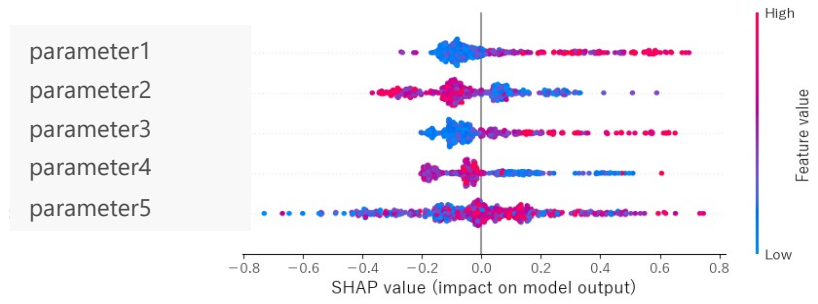


Fig. 2

※1　例えば、湿度をパラメータとするか水蒸気圧をパラメータとするかなど
※2　帰納的に情報を取り出し特徴量を抽出するが、気象要素の傾向の説明がある程度可能と考え、演繹モデルとした

以上より、演繹モデルから抽出された体調レベルを、観測された体調レベルで補正すること（fig. 3 ）で、平均的に確からしい体調レベルとして目的変数行列（6）を生成した

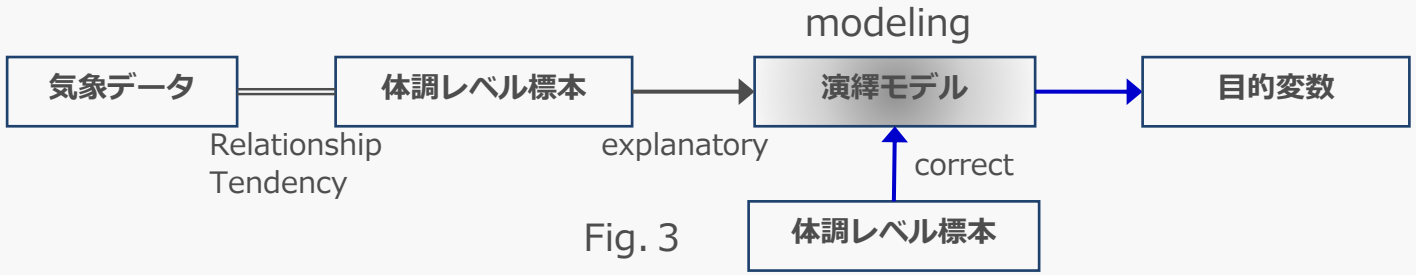


Fig. 3

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{p1} & \cdots & y_{pn} \end{pmatrix} \quad \cdots\cdots (6)$$

2．予測アルゴリズム生成

2-1.予測アルゴリズムの構成

体調予測を行う手段として、本システムでは2つのアルゴリズムを用いて実現させている。異なる2つのアルゴリズムを使うことで誤差の低減を狙った。
アルゴリズムの構成として時系列モデルを考えたが、例えば、多変量ARIMAモデルなど説明変数に定常性が求められいることから条件が合わず、また非線形で表現が可能な状態空間モデルについても、状態変数の解釈が複雑になることから、本目的に合わないことが分かった。そこで、本システムでは予測を目的とした回帰系のアルゴリズムを使用した。

2-２.予測アルゴリズムの生成

第1の予測アルゴリズムでは、予測モデルのロバスト性を上げるために目的変数を抽象化した。いわゆる2値化を課すことで、目的変数のノイズを低減させることが出来る。アルゴリズムには一般化線形モデルの改良版を用いて、正則化、クロスバリデーションを加えることで予測正答率を確保した。

一般化線形モデルの代表であるロジスティック回帰モデルを式7,8,9に示す。式7は説明変数を線形和で表した線形予測子である。最尤法でパラメータθを推定し、説明変数行列を式9に代入することで、体調レベルを確率値として導出することが出来る。

$$f(t) = \frac{\sum \exp\big(g(\theta, X)\big)}{1 + \sum \exp\big(\theta, g(X)\big)} \quad \cdots\cdots (7)$$

$$g(\theta, X) = X\theta + \varepsilon \quad \cdots\cdots (8)$$

$$\hat{p}_1 = f(\theta^T, X) \quad \cdots\cdots (9)$$

このモデルは、誤差に正規性が求められるいわゆるパラメトリックなモデルである。このモデルを使うことで、規則性を担保した予測値が導出できると考えた。

第2の予測アルゴリズムは、目的変数の連続性を効果的に取り込むべく、非線形回帰分析の考え方を用いた。このモデルも同様に、正則化、クロスバリデーションを行うことで回帰説明率の精度を上げている。また予測値の出力に対し、期待値と分散が導けるので、区間推定としての予測値判定も可能となっている。

アルゴリズムには、多項式といったモデルを事前に仮定しないノンパラメトリックモデルであるサポートベクター回帰モデルを使用した。
このモデルを第2のモデルに用いた理由は、パラメトリックな第1のモデルに対し、分布を規定しないノンパラメトリックなモデルと組み合わせることで、現実を捉え、かつ平均化した予測ができると考えた。

さらにこのモデルにカーネル関数を用いることで、事象を非線形で表すことができ、さらに分類のマージンを評価することも可能となる。式10にカーネル関数を使った回帰式を示す、ここでαはLagrange定数としている。またカーネル関数にはRBF（Radial basis function kernel)）(11)と、線形カーネル(12)を組み合わせた。

$$\hat{p}_2 = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K\left(x^j, x^i\right) + c \qquad \cdots\cdots (10)$$

$$K\left(x^j, x^i\right) = \exp\left(-\frac{\left(x^i - x^j\right)^2}{\sigma^2}\right) \qquad \cdots\cdots (11)$$

$$K\left(x^j, x^i\right) = (x^i x^j)^k \qquad \cdots\cdots (12)$$

第1の予測アルゴリズム同様に、クロスバリデーションを用いてパラメータを最適化する事で、正答率の高い体調予測が可能となる。

これら、2つのモデルの出力結果を統計的に解釈し、最終的な体調予測値を確率として提供できる

3．体調予測値の算出

3-1.気象数値予測データの収集

ここでは、10日間の体調予測方法について記載する。解析フローはfig.４に示す

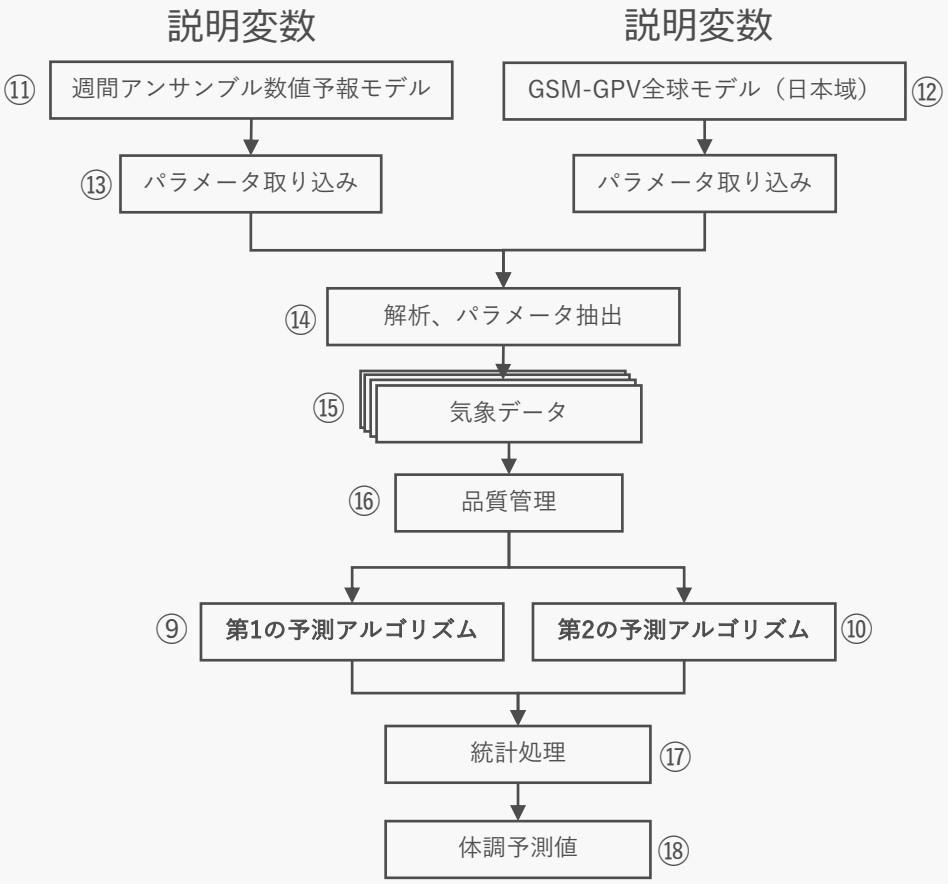
体調予測に使う気象データは、気象庁が配信する数値予報モデルのデータを用いる。使用する数値データは、日本周辺を0.5625度、51メンバーで計算された週間アンサンブル数値予報モデルGPV※３と、同じく日本域を0.1度×0.125度で計算するGSM-GPV※４の2種類の数値予報データとした。

※3 週間アンサンブル数値予報モデルＧＰＶ(日本域)：週間単位の気温、風、水蒸気等の状態について、スーパーコンピュータを用いてアンサンブル予報の手法により、3次元の格子で予測したデータ

※4 GSM-GPV：全球数値予報モデルＧＰＶ(日本域)　日本域の大気を対象に、格子間隔（水平分解能）約13kmとして、未来の気温、風、水蒸気量、日射量等の状態について、スーパーコンピュータを用いて3次元の格子で予測したデータ

数値予報モデルのデータは、Grib2という世界標準フォーマットのBinaryファイルで配信されるため、そこから気象パラメータを抽出するためプログラムを作成する必要がある。気象データのビジネス活用を推進している、気象ビジネス推進コンソーシアム（ WXBC）からプログラムをダウンロードし、本システムに合うように改良を加え用いている。

予測値算出のプロセス



番号	内容
⑪	気象庁提供の数値予報データ（週間アンサンブル）ダウンロード
⑫	気象庁提供の数値予報データ（全球モデル日本域）ダウンロード
⑬	Binaryデータを専用プログラムで取り込む
⑭	気温、湿度、気圧などのパラメータを抽出
⑮	気象データ行列生成（説明変数）
⑯	欠損値、異常値の評価と補正
⑰	予測アルゴリズムの結果の解析と予測値の統計処理
⑱	体調予測値算出

Fig. 4

数値予報データについて説明する、数値予報は、大気の状態を表す物理式（大気運動方程式、熱力学方程式、水蒸気の方程式、気体状態方程式・・・）を定められた時間間隔で積分して将来を予測するものである。
しかし微分方程式に与える初期値は誤差を含んでおり、その誤差によっては時間経過と共に解に大きく差が生じてくる。
週間アンサンブル数値予報では、その初期値に解析値より導いた誤差を積極的に印加した51メンバーを設定し、それぞれ演算を進め平均値と分散を導出している。
GSM-GPVは最も良い初期値から予測を行う手法で、決定論的予測と言われる。

週間予報の課題を記する。fig. 5 に気温のアンサンブル予測を示す。各色の線で示すのが51のメンバーである。見て取れるように時間の経過と共に分散が大きくなることが分かる。これは予測開始時点の気象状況の微妙な変化でも時間とともに大きなずれとして現れ、予測が安定しないことを意味する。

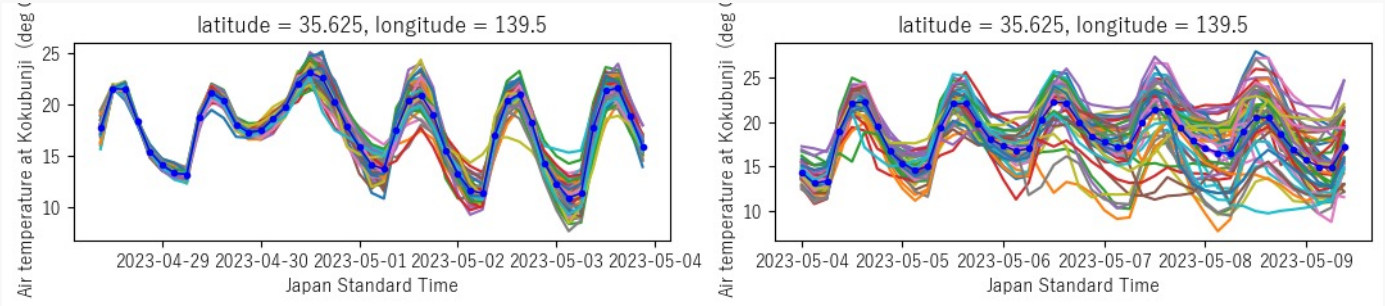


Fig. 5

そこで、本システムでは、週間アンサンブル数値予報の分散値（精度情報）を確認し、基準以下の場合は決定論的予測であるGSM-GPVを用い、基準以上では、アンサンブル数値予報の平均値をもちいて時間経過における予測値の精度を上げている。

3-2.体調予測値の算出

これらの処理を行った数値予報データから、10日間の気象予測データ行列（13）を生成する

$$\mathrm{X}^{*}=\begin{pmatrix} x_{11}^{*} & \cdots & x_{1n}^{*} \\ \vdots & \ddots & \vdots \\ x_{p1}^{*} & \cdots & x_{pn}^{*} \end{pmatrix} \quad \cdots\cdots (13)$$

最後に、上記気象予測データ(13）を、予測アルゴリズム１と２に入力し、それぞれの出力結果を解析し、重みづけ平均化(14)する事で、体調予測データとして提供することができる。

$$\tilde{p}_t = \omega_1 \hat{p}_1 + \omega_2 \hat{p}_2 \quad \cdots\cdots (14)$$

Fig.6がその最終的に求められた体調予測値である。波形が体調変化に時間的幅を与えるためにスプライン近似をかけて提供していることが特徴である

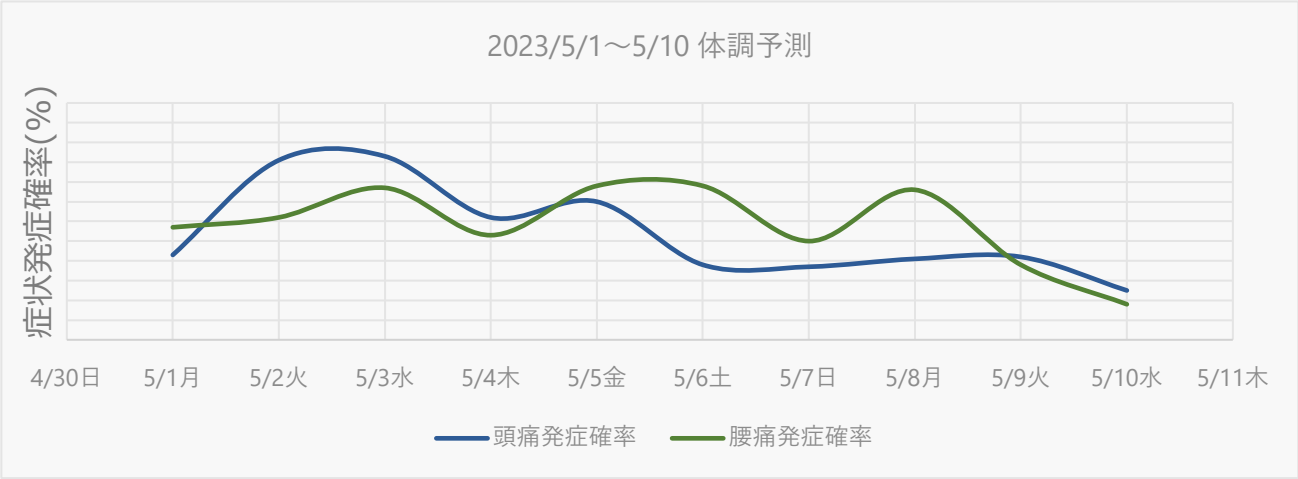


Fig. 6

４．考察と課題、謝辞

以上により、本体調予測システムでは、対象の変化に影響される気象要素に対し、単一の気象要素に特化した判断ではなく、各気象要素が相互に関係しあい、また累積性を加味することで、平均的に確からしい体調予測値を提供することができた。

しかし、予測値の正答率を100%にすることは困難である。それを生じさせる誤差の要因として2点示す、

1つ目は数値予報の精度である。特に週間数値予報は期間後半になると信頼性が大きく下がることから、対象期間の短縮が望まれる。それに伴い数値予報モデルも、GSMからMSMに置き換え、最長3日程度の予測が可能なようにシステムを変更することが望ましい。

2つ目は、モデル生成時の体調レベル（目的変数）に個人差、年齢差、性別差、があることである。あるべき姿としては個人のおかれている状況や条件を個別にサンプルしモデル生成することであるが、大きな負担となる。そこで本システムでは体調レベルの傾向を平均化することで一般的な精度は確保しているが、さらなる精度向上を価値とする場合は、予測値のカスタム化など考慮に入れる必要がある

以上、サンプルを提供いただいた小沢鍼灸院の患者様、また個別に情報を提供したいただいた方々に感謝させていいただく

今後は、本システムの価値を理解いただいた方々と共に、ビジネスに適用すべく努力をしていきたい

評価関数　$\arg min(X - X_o)^T \Sigma^{-1} (X - X_o)$　のxを求める

$y_t = g(y_t)$　　$x_{it} = x_{jt} x_{kt}$　　$x_{it} = x_{jt} - x_{kt}$

$y^\lambda = \frac{y^\lambda - 1}{\lambda} (if\ \lambda \neq 0)$　or　$\log(y_t)\ (if\ \lambda = 0)$　　$\tilde{y}_t = \sum (\omega_{1h} y\_std_t + \omega_{2h} \widehat{y\_std_t})$

$y\_std_t = \frac{y_t - \mu_y}{\sigma_y}$

T_{max}: 最高気温、T_{min}: 最低気温　　P_{ave}: 平均海面気圧、p_{min}: 最低海面気圧

RH_{ave}: 平均湿度、RH_{min}: 最低湿度　　WP_{ave}: 平均水蒸気圧　$X^* = \begin{pmatrix} x^*_{11} & \cdots & x^*_{1n} \\ \vdots & \ddots & \vdots \\ x^*_{p1} & \cdots & x^*_{pn} \end{pmatrix}$

WS_{ave}: 平均風速、WB_{mod}: 代表風向

$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$　　$x_{ti} = log x_{tj}$　　$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix}$

$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{p1} & \cdots & y_{pn} \end{pmatrix}$

$$x_{it} = x_{jt} - x_{kt}$$

$$p(y_t|b) \sim Distr\left(\mu_t, \frac{\sigma^2}{\omega_t}\right)$$

$$\tilde{y}_t = \sum(\omega_{1h} y\_std_t + \omega_{2h}\widehat{y\_std}_t)$$

リンク関数 $g(\mu_t) = X\beta + Zb + \varepsilon_t$

$$g(\theta, X) = X\theta + \varepsilon$$

$$f(t) = \frac{\sum \exp\big(g(\theta, X)\big)}{1 + \sum \exp\big(\theta, g(X)\big)}$$

$$K\big(x^j, x^i\big) = \exp\left(-\frac{\big(x^i - x^j\big)^2}{\sigma^2}\right)$$

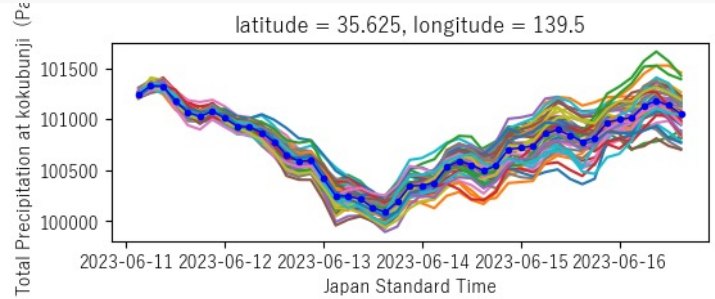
$$\hat{p} = f(\theta^T, X)$$

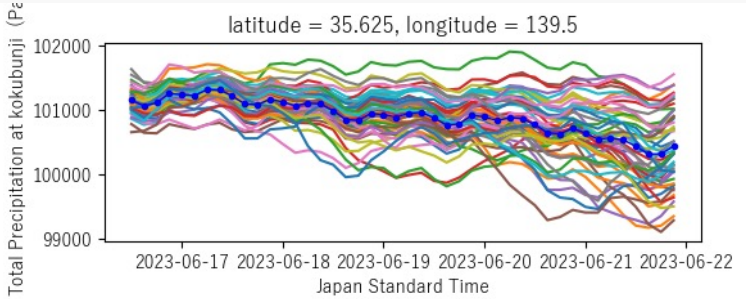
$$\arg\min \frac{1}{2}||\omega||^2 + C\sum\ h(y^i - f\big(x^i\big)) \qquad \tilde{p}_t = \omega_1\hat{p}_1 + \omega_2\hat{p}_2$$

$$f(x^i) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K\big(x^j, x^i\big) + c$$

$$K\big(x^j, x^i\big) = (x^i x^j)^k$$

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{p1} & \cdots & y_{pn} \end{pmatrix}$$


latitude = 35.625, longitude = 139.5
Total Precipitation at kokubunji (Pa
101500
101000
100500
100000
2023-06-11 2023-06-12 2023-06-13 2023-06-14 2023-06-15 2023-06-16
Japan Standard Time


latitude = 35.625, longitude = 139.5
Total Precipitation at kokubunji (Pa
102000
101000
100000
99000
2023-06-17 2023-06-18 2023-06-19 2023-06-20 2023-06-21 2023-06-22
Japan Standard Time