

多様な地球環境データを対象としたマイニングシステムの構築

生駒 栄司^{†,†††} アーマド シャーキール^{††} 小池 俊雄^{††,†} 喜連川 優^{†††,†}

[†] 東京大学地球観測データ統融合連携研究機構

^{††} 東京大学大学院工学系研究科

^{†††} 東京大学生産技術研究所

1 はじめに

昨今の地球観測技術の進歩により、取得される地球環境データの量は爆発的に増大しつつある一方、これらのデータを利用する研究者が処理出来るデータ量がその増加量に追いつかず、貴重なデータが十分利用されず眠っているのが現状である。

そこで本研究では、当該分野の研究者と密接に協力し、データマイニング技術を用いて膨大かつ多様な地球環境データを融合し、さまざまな時系列処理も含めて解析するツール群の開発を行い、非計算機分野の研究者にも容易に操作可能なインターフェースを実装したシステムを構築することで、従来手法では解析が困難であった多様な自然現象解明のための支援システムを構築した。

2 自然現象の相関解析

気象学においては従来より統計解析による現象間の関連抽出手法が用いられてきたが、その中でも最も一般的なものの一つが相関係数を用いたものである。Wallace[1]らは北緯45度、東経165度の格子点における500hPa面の高さの変動を基準として相関係数の分布を調べた結果、北太平洋とアメリカ中部から東岸に正の相関域、アメリカ西岸から中部にかけて負の相関域が存在することが判明した。これは、基準となる格子点の気圧が高くなると、正（負）の相関域では気圧が高く（低く）なることを示している。このような相関係数の分布はPNAテレコネクションパターン（PNA:Pacific/North America）と呼ばれ、この現象は現在では力学的には赤道域における熱源により励起された定在ロスビー波として解釈されており、相関解析による発見が理論的にも説明されている。

Development of Data Mining System for Huge Variety of Earth Environmental Data

Eiji IKOMA^{†,†††} Shakil AHMAD^{††} Toshio KOIKE^{††,†} Masaru KITSUREGAWA^{†††,†}

[†]EDITORIA, The University of Tokyo

^{††}Graduate School of Engineering, The University of Tokyo

^{†††}Institute of Industrial Science, The University of Tokyo

このように、旧来の物理的理論から現象発見へとアプローチする手法とは逆に、まず現象間の関連性を発見しておいてからそれを物理的な理論に従って説明あるいは解釈していくという方向も自然現象解析のための一手法であると考えられる。

3 相関解析ツールの開発とシステム構築

2に示したように、気象学の分野において統計解析的な手法で現象間の関連抽出が行われてきた例はあるが、本研究で想定するような大規模かつ多種多様なデータを対象とする場合、上述のような汎用ソフトウェアで処理するのは非常に困難である。また、本研究で対象としているデータは、単純な2データの相関計算にとどまらず、時間・空間・変量を超えた相関性を持つ可能性があるため、柔軟な条件指定で大容量データを扱うことが可能なツールが必要である。

そこで本研究では、東京大学地球観測データ統融合連携研究機構で運用を行っているDIAS(Data Integration and Analysis System)[2]のデータ統合・情報融合コアシステム（DIASコアシステム）と連携し、次節に示す各種再解析データを対象とした時間・空間・変量に加え時間解像度や空間解像度などを柔軟に指定して相関解析が可能なツールを開発した。

3.1 対象データ

現在、DIASコアシステムに収蔵されている(1)Geopotential height, (2)Air Temperature, (3)Specific humidity, (4)Zonal wind, (5)Meridional wind, (6)Pressure vertical velocity, (7)Outgoing Longwave Radiation, (8)SST and sea ice surface temperature, (9)Upward long wave radiation flux at surfaceの各データが利用可能であり、いずれも1979-2004年、(1)~(6)は200/500/850hPaの各気圧面ごと、(6)はNCEP[3]再解析データ、それ以外はJRA25[4]長期再解析データ、(1)~(6)は1.25度グリッド、(7)~(9)は2.5度グリッドの日毎データである。

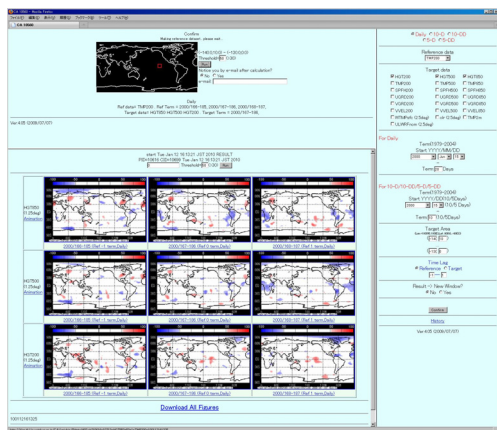


図 1: 相関解析システムのユーザインターフェース

3.2 システム構成

本システムにおいて、ユーザが Web インターフェース上で指定した条件や処理内容は Web サーバを經由して DIAS サーバ内のマイニング処理部およびデータ視覚化処理部に送られる。その条件に基づいてマイニング処理部は SQL を生成し、データアーカイブシステムのデータベースにデータ要求を送出、そこで得られたデータはマイニング処理部で処理内容に基づいた相関解析計算が行われ、結果がデータ視覚化処理部に送られる。そこで最初に指定した視覚化条件に基づいた視覚化処理が行われ、その結果は Web ブラウザ上に表示、ユーザが結果の確認を行う。

3.3 解析処理の流れ

本システムは図 1 に示すインターフェースを有しており、この Web ページ上で、ユーザは左側フレームにおいて (1) 相関解析の比較元となるデータ (1 データ)、(2) 比較対象データ (複数選択可能) (3) 比較年月日と期間、(4) 比較元となるエリア、(5) タイムラグ、(6) 結果表示ウィンドウ指定 (別ウィンドウか、同一ウィンドウ内か) を指定する。すると、左上フレームに世界地図上で解析対象として指定したエリアが表示され、同時に視覚化表示時の閾値設定および処理モード選択 (即座に実行し結果を表示するか、計算後にメールで通知し結果を確認するか) の指定を求められ、ここで実行ボタンを押すと解析処理が実行される。

結果は左下フレームに表示され、横方向はラグ日数の異なる結果が並び、縦方向は異なる変数を示している。すなわち、1つの結果データは、比較元データの既定領域の領域平均時系列データと、ある比較先データの、比較元とのある時間ラグのある年月日の、全球全点における相関解析計算の結果を示しており、指定

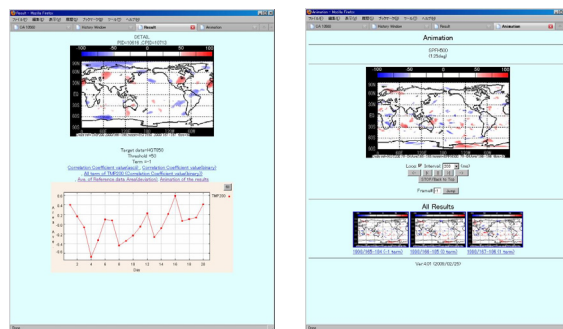


図 2: 解析結果の利用 (詳細表示・アニメーション)

した閾値に従って正相関は赤、負相関は青のそれぞれ濃淡で大きさを示している。従って、横方向に結果を比較することで、時間の推移による相関性の変化 (どの程度の時間ラグがあって影響を及ぼしたか) が解析でき、縦方向の対比で比較元データがどのデータと相関性が高いかの把握が可能となる。

また、この結果フレームにおいて関心のあるデータをクリックすることにより、図 2 に示すように時系列グラフ等より詳細な結果の確認、ダウンロード、アニメーションによる時系列変化の確認が可能である。

4 今後の予定

現在、本システムは DIAS プロジェクトにおいて各分野の専門家によって利用され、自然現象解析の新たなアプローチの 1 つとして数々の知見を創出している。今後も継続的に利用者と協力しながら、さらに多様なデータへの対応、より高度な解析アルゴリズムの導入、より柔軟な利用が可能なインターフェースの構築を続けていく予定である。

謝辞

本研究は、文部科学省による委託研究費「データ統合・解析システム」からの支援を受けて行われた。よってここに謝意を表します。

参考文献

- [1] Wallace, J.M. and Gutzler D.S., "Teleconnections in the geopotential height field during the Northern hemisphere winter", Mon. Wea. Rev., vol.109, pp.785-812, 1981.
- [2] DIAS, <http://www.editoria.u-tokyo.ac.jp/dias/>
- [3] NCEP Reanalysis, <http://www.cdc.noaa.gov/>
- [4] JRA25 Data, <http://jra.kishou.go.jp/JRA-25/>