プロジェクト名：　データ同化による複雑システムの定量的理解と計測デザイン

プロジェクトディレクター：　中野 純司　教授（統計数理研究所）

〔1〕 研究計画・研究内容について

(1) 目的・目標

広範囲な分野の複雑なシステムを対象とする研究の具体的な問題解決に向けた共同研究をとおして、シミュレーションとデータ解析の作業を一体化する手法の高度化と一般化を実現する。あわせて、未適用分野の発掘を戦略的にすすめ、個別科学を横断的につなぐ新しい学問領域を創る。

(2) 必要性・重要性（緊急性）

コンピュータの計算能力の向上とともに大規模かつ複雑精緻なものとなるシミュレーションモデルと、飛躍的に増大することが予想される観測・計測データの融合技術の開発は急務である。あらゆる研究分野で、シミュレーションが生み出す大量の計算結果をどのように評価するのか、また研究対象に関する多面的網羅的な観測・計測データをシミュレーションモデルの改良にどう生かしたらいいのか、エキスパートがごくかぎられた部分を不完全に考察しているのが現状である。このように、シミュレーションと大量データを連係させる統計数理をないがしろにしてきたことが、あらゆる科学研究の推進のボトルネックとなってきたと言えよう。データ同化はまさにこの問題に真正面から取り組む研究で、その研究の端緒はまだ 20 年に満たない。そのこともあり、日本においてはいまだ統計数理研究所以外に、手法の総合的研究を行っているところがないという、まことに危機的状況である。データ同化技術は、初期投資として高性能の計算機を用意しそれを有効利用することで、コストを下げつつ一方推定精度は向上させる観測・実験システムの提案も可能であり、大規模予算を投入する大型装置や大規模実験の設計においても今後必須の道具となることは確実である。

(3) 期待される成果等（学問的効果、社会的効果、改善効果等）

学問的効果としては、理論を基礎とする計算推論技術であるシミュレーションと、実験・観測データを基盤とする推測手段であるデータ（統計）解析法の両者を統合する新しい学問体系の構築がまず挙げられる。このようなシミュレーションと大量データ解析を両輪とする学問体系の確立は各個別科学分野において長年の夢であり、その波及効果はシミュレーションを研究手段に用いているあらゆる分野に及ぶ。本プロジェクトは、人間とコンピュータが協調してつくりあげる“データにもとづいて考えるスーパー科学脳”の実現を目指していると言える。

データ同化の分かりやすい目的に、予報を行うための最適な初期条件の探索や、スケールが全く異なる現象どうしをシームレスに連結するシミュレーションモデル内のパラメータ設定がある。これらにより、実はすぐにでも手元のシミュレーションを高性能化することが可能であるにもかかわらず、具体的に限定された計算資源の中でどう工夫したらよいか暗中模索の状態となっている。仮想観測ネットワーク実験や感度解析が可能になるのもデータ同化の恩恵の一つである。社会的ニーズに十分応えるスピードで、さらに限定された研究費内で最大限の知見を獲得する、観測や実験システムの立案が可能となる。時間・経費を節約できる効率的な実験・観測システムを構築することは、納税者である国民の理解を得る上で非常に大切な観点である。

これまでは、シミュレーションのような順問題的に課題を解決する研究者の養成が学術分野の体制であったが、本プロジェクトの推進により、データから理論に遡る、まさに逆問題解決のセンスも兼ね備

えた新しいタイプの研究者の養成が可能になる。これにより、どのようにして取り扱ったらよいか分からない問題に果敢に取り組める人材の育成プログラムを強化できる。

(4) 独創性・新規性等

データ同化の計算基盤を与えるのが状態空間モデルにもとづくフィルタリング計算技術で、その研究開発では統計数理研は誇るべき経験と実績がある。研究所内に設置された、予測発見戦略研究センターデータ同化グループは、先端的同化手法やその上位概念物であるメタシミュレーション法の研究開発とともに、インパクトのある新規適用分野の開拓を行ってきた。実際に、津波、海洋潮汐、宇宙空間と、データ同化の概念を比較的柔軟に理解してもらえる分野で新しい研究テーマを複数開発できたばかりでなく、ゲノム情報分野、“ものづくり”においてデータ同化の研究を開始し、まさに「統計数理研究所データ同化グループが種をまき双葉にまで育て上げた」と言える。データ同化の公開講座を同グループメンバーで平成 20 年度（4 日間連続）と 21 年度（2 日間連続）に開講したが、受講希望人数が多く締め切り前に早々と応募を締め切るなど、非常に好評であった。また現在“データ同化”のキーワードでもってインターネットで検索すると、トップサイトの 3 つは本プロジェクトに直接関連したページである。これらから、データ同化手法の研究を本グループが先導していることを物語るとともに、新規適用分野の開発を含めてデータ同化の総合的研究推進において国内の中心的役割を果たしていることは明らかである。

(5) これまでの取り組み内容の概要及び実績

統計数理研究所・予測発見戦略研究センターの平成 15 年 9 月の創設時は、データ同化グループは動的磁気圏モデル研究グループという名を掲げ、大量データにもとづく地球磁気圏の時空間経験モデルの構築を主たる研究目的としていた。平成 16 年 10 月に JST の CREST「シミュレーション技術の革新と実用化基盤の構築」研究領域に、「先端的データ同化手法と適応型シミュレーションの研究」題目で、樋口教授を代表とする同グループメンバーが提案する研究プロジェクトが採択されたため、そのプロジェクトが実質的に動き始めた平成 17 年 4 月にグループ名をデータ同化グループと変更し、研究推進体制を整備した。これまで所長のリーダシップにより、同グループに研究スペースの配分が優遇されおり、また最低限の基盤的研究費の継続的な支援が行われている。なお、その CREST プロジェクトは平成 21 年度に終了するが、中間評価では高い評価を得た。

機構内におけるデータ同化の研究は、これまで主に JST/CREST 事業の他には、次世代スパコンの研究開発資金により、理化学研究所を中心に機構外の複数の国内の研究機関に所属する多くの研究者群を巻き込んで行われているアプリケーションソフト開発プロジェクト(「次世代生命体統合シミュレーションソフトウェアの研究開発」）内で、生命体データ同化技術とアプリケーションの開発が行われている。その開発も統計数理研究所が担当している。これに対して、第 1 期の新領域融合研究センターの活動としてのデータ同化の研究は統数研との共同研究が、融合プロジェクト「機能と帰納」により国立極地研究所と、また育成融合プロジェクト「デジタル細胞を利用した仮説発見」により国立遺伝学研究所と小規模に行われているのが現状である。

(6) 国内外における関連分野の学術研究の動向

データ同化にはオンライン型（逐次データ同化と呼ばれる）とオフライン型（非逐次データ同化）がある。非逐次型データ同化は比較的昔（1990 年代中頃）から欧米の気象・海洋学の現業機関を中心に研究開発が進んできており、国内においては JAMSTEC と気象庁がその中心的存在である。一方、統計数理研究所のデータ同化グループは、オンライン型である逐次データ同化手法の研究とその応用を行って

いる。同グループは、データ同化手法を統計科学の枠組みで正確に定義し、統計科学や情報科学の分野において蓄積されてきたアルゴリズムやモデリングに関する研究成果を利用しながら、これからの計算機インフラを視野に入れつつ、実装が平易かつ広い分野に適用できる逐次データ同化手法の開発を行ってきた。このような研究を行っているのは国内では同グループのみであったが、気象庁にて逐次データ同化手法の一つであるアンサンブルカルマンフィルタを用いた予報の現業化が始められつつあることからしても、逐次データ同化技術の有効性は国内においてはようやく認められてきたと言えよう。また、大気海洋結合モデルへのアンサンブルカルマンフィルタの適用は同グループが世界初であったが、同様のデータ同化研究は、中国科学院とアメリカ大気研究センター（NCAR）でも進められている。

〔2〕 研究計画

(1) 全体計画

本プロジェクトにおいては、数理・計算、モデリング、データデザインの 3 チームを構成し、それらが有機的に協力して研究を進め、目的の達成に努める。運営の取りまとめや研究の統合化は数理・計算チームの主たるメンバーが所属する統計数理研究所のデータ同化研究開発センターが行う。モデリングチーム、データデザインチームはそれぞれ極域科学、生命科学における具体的で明確な現象を扱うことが特徴であるが、数理・計算チームでもデータ同化の理論のみならず具体的な現象も扱う。

全体を通して不断の、研究体制の編成、情報収集・整理、研究会・ワークショップの開催、研究体制の見直し、を行う。

(2) 各年度の計画

平成 25 年度

数理・計算チームでは、階層的シミュレーションへのデータ同化適用事例として、地球物理シミュレーションと携帯電話のようなセンサ情報に代表されるサイバーフィジカルシステムとを統合する地震津波災害情報統合システムや、市町村内におけるミクロシミュレーションと、地域間や海外を考慮したマクロシミュレーションを統合した感染症予報システムの開発に取り組む。これらの研究では、まずは、データのアクセシビリティを調査することから始める。今年度は、被災者の位置情報、定点観測医療機関ごとの感染者数等についてのその調査を行い、小規模シミュレーションによるプロトタイプを作成する。

モデリングチームは、グローバル MHD シミュレーションモデルへのデータ同化に使われる状態空間モデルや誤差について、モデルのパラメータと誤差の重みづけを再検討する。初期実験結果と観測データとの比較により同化システムの改善を目指す。また、大気レーダーの観測においてこれまで推定の難しかった微速の鉛直風推定手法について適応的信号処理を用いた手法を開発し、シミュレーションなどを用いてその推定精度を示した。国内の大型大気レーダー（MU レーダー）を用いた実験データを取得しており、これに基づき精度の実証を試みる。

データデザインチームは、前年度のデータ同化実験の進め方の検討結果にもとづき、データ獲得（観測）を大規模に行う。さらには同年度までの解析結果を小括し、対外発表などを通じて他の系への応用可能性を探る。

平成 26 年度

数理・計算チームでは、感染症予報システムの開発に加え、宇宙ごみ（デブリ）の空間分布の特異変動モデリングに本格的に取り組む。感染症予測システムでは、感染者数の時空間分布から感染伝播における地域間の結び付きを推定し、ひとつの都市を対象とするシミュレータを全国レベルのシミュレータ

に統合する。デブリ空間分布のモデリングでは、膨大な公開観測データから逆問題的方法により、デブリの空間分布の特異変動の原因となる事象を検出する研究に着手する。時系列画像中から常に暗く写るデブリ（微光デブリ）をベイズ的手法により効率良く検出するアルゴリズムの開発に着手する。さらに、親物体とその破片の軌道の公開データから破砕イベントで発生する破片の方向分布の正確なモデル化に取り組む。

モデリングチームは、グローバル MHD シミュレーションから再現される電離層のプラズマ対流や沿磁力線電流の分布について観測値に近づくように境界パラメータサーベイを行う。具体的には、シミュレーション内で 10 分程時間を進めた後の計算値が観測値に近づくように、合理的な範囲で内部境界パラメータを変えてシミュレーションを行い、もっとも観測値に近い結果が出るパラメータを推定する。また、大気レーダーデータ解析に関しては、レーダー観測可能な空間スケールの下限をさらに下回るスケールの乱流ダイナミクスについて､状態空間モデルを用いて高精度に推定する手法について研究する。観測データと物理量が領域積分系観測モデルを経て結合されるが、積分領域内のローカルな時間発展が大きい問題であるため、この逆問題を効率的に解くスキームについて開発を行う。

データデザインチームは、前年度からの大規模なデータ獲得を引き続き行うとともに、数理・計算チームと合同でデータ同化実験結果の多面的かつ慎重な検討を行い、データ同化実験用のデータセットとしての完成度を高めることに注力する。

平成 27 年度

数理・計算チームでは、感染症予報システムに関しては、開発した統合シミュレータに実際にデータ同化を行い、リアルタイムな予報が可能かどうかを検証する。また、インフルエンザなどの周期的な感染流行は、ウィルスの進化と出生・加齢・死亡による対象集団の免疫構造の変化を反映していることを受けて、遺伝情報や血清調査などの関連するデータの利用可能性を調査する。デブリ空間分布のモデリングでは、その特異変動と原因事象の候補との間の相関性を評価する手法を開発する。また、微光デブリ検出アルゴリズムを実装したソフトウェアを複数の観測所に導入して、地上光学観測を通じた実証試験を行い、実運用上の問題点を洗い出す。

モデリングチームは、グローバル MHD シミュレーションモデルに同化システムを実装し運用する。具体的には、計算値と観測値を客観的に比較、インデックス化し、評価関数を構築し、評価関数の分布にしたがって内部境界パラメータの最適値を自動的に算出する。同化システムを運用した結果得られるより現実に近い計算値と、非同化システムによる従来の計算値について 3D 可視化ツールによる比較を行う。また、大気レーダーデータ解析に関しては、乱流スケールの推定問題について研究を進めるとともに、PANSY レーダーを用いた高次物理量観測を進める。これまでに得られたソフトウェアを観測システムの一部として組み込み、高度な情報を安定的に国内へ輸送するとともに、本課題で得たノウハウについて関連する他の研究課題を含む幅広い分野への水平展開を検討する。

データデザインチームは、成果公開を通じてデータ同化手法の生命科学分野、特に定量イメージングデータを利用した発生細胞生物分野への普及・発展に努める。

〔3〕 研究推進・実施体制

・研究代表者

〔統計数理研究所〕　中野純司

・共同研究者

〔統計数理研究所〕　樋口知之、上野玄太、吉田　亮、中野慎也、長尾大道、齋藤正也、上津原正彦、鈴木香寿恵

〔国立極地研究所〕　宮岡　宏、門倉　昭、小川泰信、田中良昌、平木康隆、中村卓司、堤　雅基、冨川喜弘、山内　恭、山岸久雄
〔国立情報学研究所〕　三浦謙一
〔国立遺伝学研究所〕　相賀裕美子、木村　暁
〔情報・システム研究機構〕　才田聡子、高橋久尚、西村耕司、二宮洋一郎、荒井律子、近藤　興
〔東京大学〕　佐藤整尚、佐藤　薫、宮野　悟
〔情報通信研究機構〕　村田健史
〔Johns Hopkins University〕　大谷晋一
〔九州大学〕　田中高史、河野英昭、吉川顕正、渡辺正和
〔京都大学〕　海老原祐輔、佐藤　亨、村上　章
〔名古屋大学〕　藤井良一、三好由純
〔Finnish Meteorological Institute〕　中溝　葵
〔千葉大学〕　中田裕之
〔気象大学校〕　藤田　茂
〔山口大学〕　松野浩嗣

〔4〕 研究の進捗状況

＜数理・計算チーム＞

(1) アンサンブル変換カルマンフィルタを用いたプラズマ圏密度分布パラメータの推定

地球磁気圏のプラズマは、地上高度 2 万 km 程度までの比較的高度の低い領域に集中して分布しており、この領域をプラズマ圏と呼ぶ。プラズマ圏のプラズマを観測する手段には様々なものがあるが、中でもプラズマ中のヘリウムイオン(He+)から発せられる極端紫外光を撮像した画像データは、プラズマ圏の大域的な描像が得られる有用なデータの一つである。図 1 は米国の人工衛星 IMAGE によって得られた極端紫外光画像データの一例で、図 2 に示すように、衛星がプラズマ圏の外に位置しているときに得られたものである。図 1 の中央に赤い円で示しているのが地球で、それより外側の発光している領域がプラズマ圏に相当する。我々は、逐次データ同化の代表的手法の一つであるアンサンブル変換カルマンフィルタを用いて、IMAGE 衛星の極端紫外光画像からプラズマ圏の He+密度分布の時間発展を推定する手法の開発を進めている。

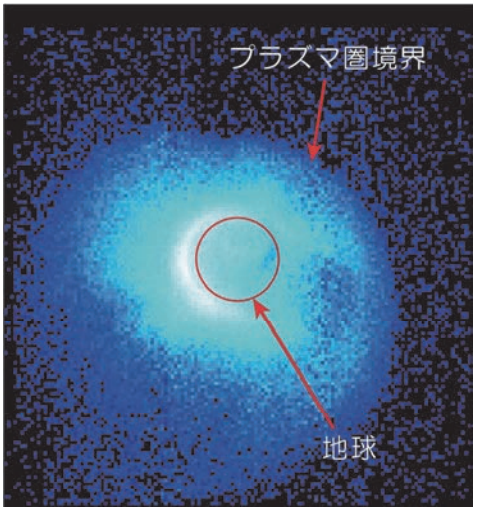

図1： 人工衛星 IMAGE によって撮像されたプラズマ圏 He+の極端紫外光画像。

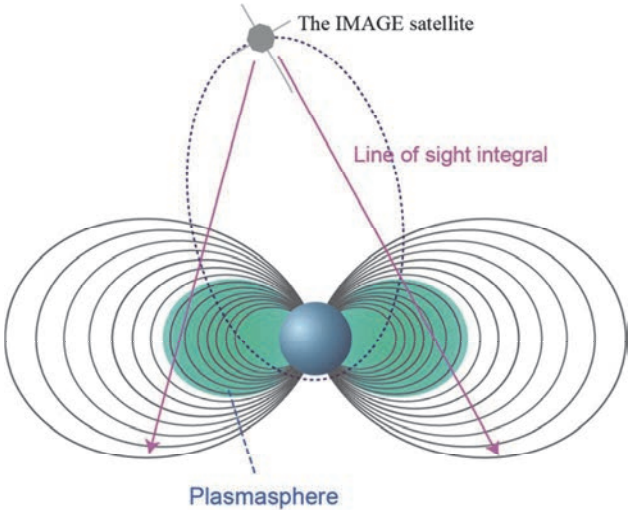

図2： 人工衛星 IMAGE によるプラズマ圏プラズマ撮像観測の概念図。

極端紫外光の 2 次元画像の情報から、プラズマ圏の 3 次元的な He+密度分布を推定するためには、密度分布構造に何らかの仮定を置く必要がある。本研究では、He+などの荷電粒子が、地球磁場の磁力線の方向に拡散しやすいという性質を考慮し、点 $\boldsymbol{r}$ における He+の密度 n が

$$n(\boldsymbol{r}) = n_{\mathrm{eq}}(\rho)\left(\frac{r_{\mathrm{eq}}}{r}\right)^{\alpha}$$

のように表されると仮定している。ただし、n_{eq} は、点 r を通る磁力線が赤道面と交わる点 ρ での He+密度であり、また $r = |\boldsymbol{r}|$, $r_{\mathrm{eq}} = |\boldsymbol{\rho}|$ である。この仮定は、プラズマ圏のモデリングで広く用いられているが、基本的に磁力線上の密度分布はよくわかっていないため、指数 α をいかにして適切に設定するかが問題であった。

そこで本年度は、最適な指数 α を決定する手法の開発を行った。1 枚の 2 次元画像から 3 次元的な構造を推定するのは一般に困難である。しかし、個々の極端紫外光画像は衛星の移動に伴って少しずつ異なる場所から撮影されたものなので、複数の画像を組み合わせることにより、3 次元的な情報を得ることができる。衛星が移動する間にプラズマ圏の He+分布も変動してしまうという問題があるが、He+分布の変動と衛星の移動の両方を考慮したモデリングを行うことで、He+の赤道面上の密度分布の変動と、磁力線方向の分布を与えるパラメータ α を同時に推定することができるようになった。パラメータ推定においては、パラメータ α の尤度を求め、これを最大化するというアプローチを用いた。パラメータ α の尤度は、

$$p(\boldsymbol{y}_{1:K} \mid \alpha) = p(\boldsymbol{y}_K \mid \boldsymbol{y}_{1:K-1}, \alpha) \cdots p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \alpha) p(\boldsymbol{y}_1 \mid \alpha)$$

となるが、ガウス分布を仮定すると、対数尤度が

$$\log p(\boldsymbol{y}_{1:K} \mid \alpha) \propto -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{y}_k - \mathsf{H}_k \boldsymbol{x}_{k|k-1})^T (\mathsf{H}_k \mathsf{V}_{k|k-1} \mathsf{H}_k^T + \mathsf{R}_k)^{-1} (\boldsymbol{y}_k - \mathsf{H}_k \boldsymbol{x}_{k|k-1}) - \frac{1}{2}\sum_{k=1}^{K}\log|\mathsf{H}_k \mathsf{V}_{k|k-1} \mathsf{H}_k^T + \mathsf{R}_k|$$

$$\propto -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{y}_k - \mathsf{H}_k \boldsymbol{x}_{k|k-1})^T \mathsf{R}_k^{-1} (\boldsymbol{y}_k - \mathsf{H}_k \boldsymbol{x}_{k|k}) - \frac{1}{2}\sum_{k=1}^{K}\log|\mathsf{R}_k||\mathsf{I} + \Lambda_k|$$

と変形できる。ただし、I は単位行列であり、Λ_k は $\mathsf{X}_{k|k-1}\mathsf{X}_{k|k-1}^T = \mathsf{V}_{k|k-1}$ として、

$$\mathsf{X}_{k|k-1}^T \mathsf{H}_k^T \mathsf{R}_k^{-1} \mathsf{H}_k \mathsf{X}_{k|k-1} = \mathsf{U}_k \Lambda_k \mathsf{U}_k^T$$

という固有値分解を行うことで得られる対角行列である。この固有値分解はアンサンブル変換カルマンフィルタの手続き中で行われるので、行列 Λ_k はフィルタの手続きの段階で既に得られていることになる。したがって、パラメータ α の対数尤度は極めて容易に求めることができる。図 3 は、ある設定で実行したシミュレーション結果をもとに、α=2 を仮定して生成した人工データから、様々な α に対して対数尤度を求めたものであるが、α が 2 のところで対数尤度が最大になっており、尤度を最大化することによって得られた α が適切な推定値になることを示している。図 4 は、実際に 2001 年 6 月 20 日のデータに対して α を推定した結果である。この例では、α がおよそ 2.1 と推定されている。また、この α の推定に用いた極端紫外光画像データを図 5 に示しておく。なお、ここでは 4 枚の画像しか示していないが、α の推定には 0900UT から 1500UT までの 6 時間に約 30 分おきに得られた 12 枚の画像データを用いている。さらに、その画像データから α=2.1 のもとで推定された赤道面上の He+密度分布を図 6 に示す。図 6 において、カラーコードで示しているのが、推定された He+密度の磁気赤道面上での値であり、白い線は同時に推定された電場ポテンシャルの等電位線である。図で地球の上側に相当する地方時 6 時の付近に強い西向きの電場が形成され、それに伴ってプラズマ圏の地方時 8 時付近の部分が浸食

される様子が推定されている。

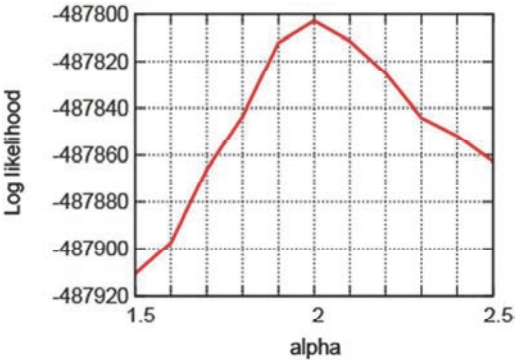

図3：ある人工データのもとでの、パラメータαと対数尤度との関係。

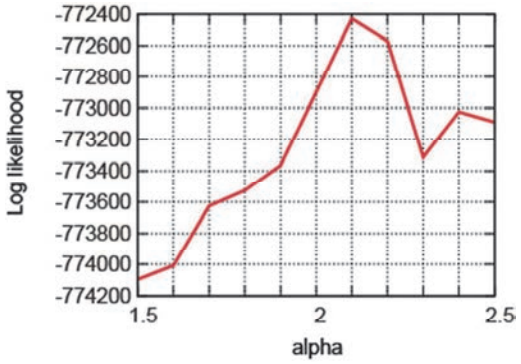

図4：2001 年 6 月 20 日のデータのもとでの、パラメータαと対数尤度との関係。

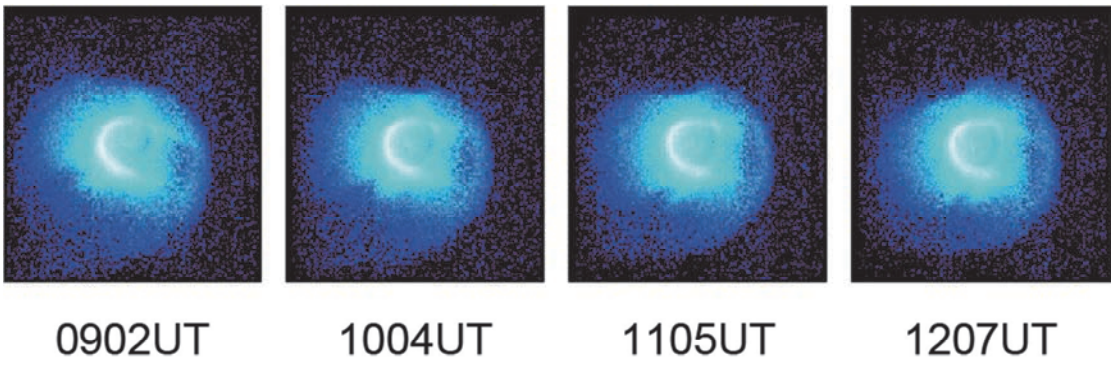

図5：2001 年 6 月 20 日に得られた IMAGE の極端紫外光画像データ。

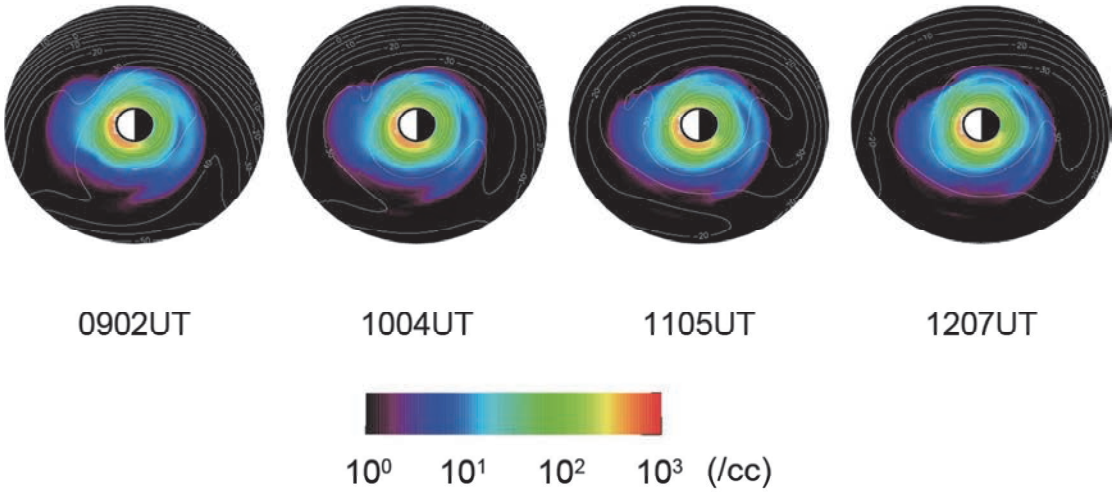

図6：IMAGE の極端紫外光画像データから、α=2.1 の設定もとで推定した 2001 年 6 月 20 日の磁気赤道面上の He+密度分布および電場ポテンシャル分布。

(2) アンサンブル変換カルマンフィルタと importance sampling の混成アルゴリズム

データ同化を行うための手法には様々なものがあるが、最近では、前進方向の計算と後進方向の計算を繰り返してデータ時系列全体に推定値を合わせる 4 次元変分法と、多数の前進方向の計算を並行して実行するアンサンブル型の逐次アルゴリズムの 2 つが代表的な方法になっている。中でも後者のアンサ

ンブル型アルゴリズムは実装が容易なために広く使われるようになってきている。多くのアンサンブル型アルゴリズムでは、状態の確率分布を多数の粒子で表現し、個々の粒子についてシミュレーションモデルを前進方向に実行することで、不確実性の情報を次の時間ステップに伝搬させる。通常、このときの粒子は分布から無作為抽出したサンプルとして扱われ、多数の粒子を使えば、大数の法則によって求めるべき確率分布の情報が精度よく表現できるものとしてアルゴリズムが設計される。この考え方は、低次元問題に対して充分な数の粒子を使う限りは有効であり、実際、そのような場合には妥当な結果が得られる。しかし、実際にデータ同化で扱われる大規模な問題においては、計算資源の制約の問題により、状態ベクトルの次元よりもはるかに少ない数の粒子しか使えないことが多く、その場合、大数の法則に依拠した考え方をそのまま適用するのが適切でない場合も少なくない。

以前より我々は、粒子数が少ない場合にも有効な方法として、確率分布をシンプレックスで表現するアプローチに着目し、このアプローチに基づいたアンサンブル平方根フィルタと呼ばれるクラスに属する手法の活用を進めてきた。プラズマ圏の He+分布推定に用いたアンサンブル変換カルマンフィルタも、このアンサンブル平方根フィルタに属する手法の一つであり、上述の結果はシンプレックス表現に基づく手法を活用した成果の一つと言える。しかし、シンプレックス表現を使った場合、確率分布の 1 次、2 次のモーメントしか考慮されないため、分布の非ガウス的な構造を扱うのが難しいという問題がある。特に、観測に非線型性や非ガウス性が含まれる場合、得られたデータの情報をうまく推定に活用するのが必ずしも簡単ではなかった。

この問題に対処するために、予測分布の計算に十分な数の粒子を使うことができない状況においても、事後分布の計算のみに大量の粒子を用い、importance sampling によって、分布の非ガウス性を表現するアルゴリズムを提案した。Importance sampling の前段階の処理として、アンサンブル変換カルマンフィルタを用い、アンサンブル変換カルマンフィルタによって得られた結果を importance sampling の提案分布として使うことで、計算の効率化を実現した。図 7 は、提案アルゴリズムの概念図である。図 7 では、2 次元的な分布が赤い点で示す粒子で表現している状況を表している。2 次元空間のシンプレックスは三角形なので 2 次元空間の分布は 3 つの点で表される。アンサンブル変換カルマンフィルタを適用すると、図の左列のようにシンプレックスが更新される。しかし、シンプレックスでは観測が非ガウス的であった場合にそれをうまく扱うことができない。そこで、図中央に示すように、アンサンブル変換カルマンフィルタによって得られたシンプレックスに対応するガウス分布から大量の粒子を抽出し、importance sampling によって観測の非ガウス性を考慮した事後分布を求める。しかし、この大量の粒子をそのまますべて次の時間ステップの予測に用いると計算量が莫大なものになってしまう。そこで、非ガウス的な事後分布の 2 次までのモーメントを表現するシンプレックス近似を生成し、このシンプレックスを構成する少数の粒子のそれぞれについてシミュレーションを実行することで、次の時間ステップの不確実性を評価する。事後分布をシンプレックスで近似することにより、分布の非ガウス性に関する情報は失われてしまうが、事後分布の計算自体には観測の非ガウス性も考慮され、非ガウス的な観測モデルをガウス分布で近似した場合よりもよい推定ができる。

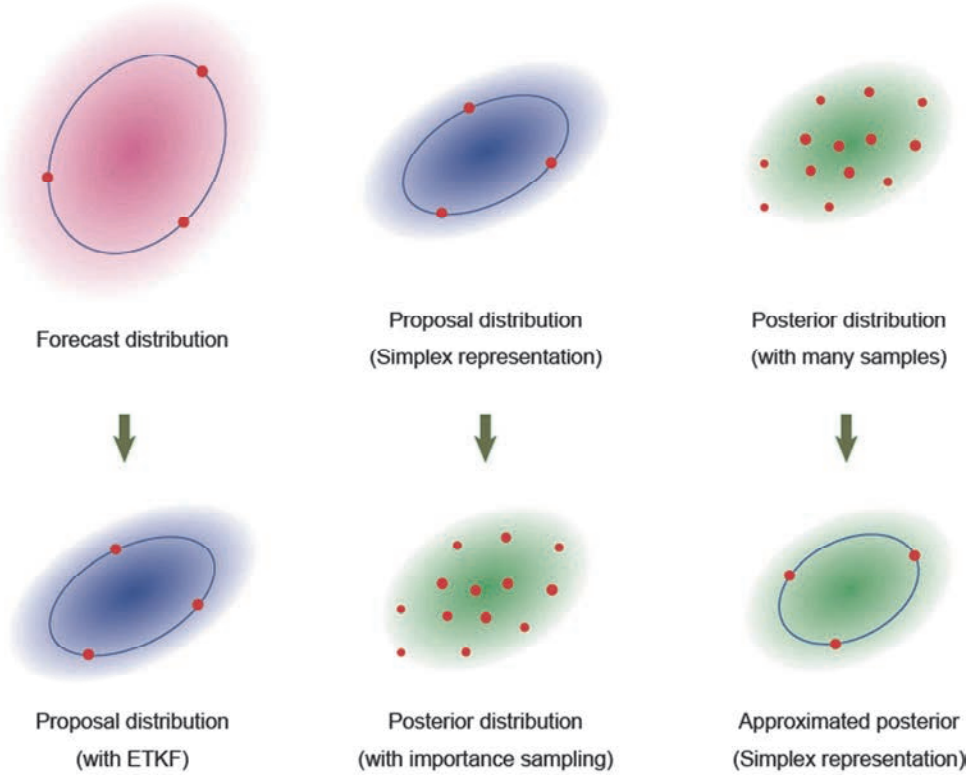

図7： アンサンブル変換カルマンフィルタと importance sampling の混成手法の概念図。

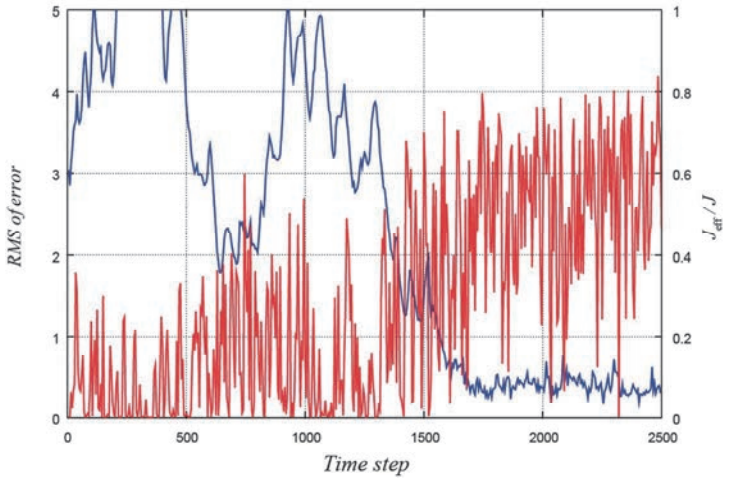

図8： 全粒子数に対する有効粒子数の比（赤）；推定値と真値との差の標準偏差（青）。

ただし、このアルゴリズムでは、importance sampling を行う際にどのくらいの数の粒子を使えばよいのかが問題となる。実際には、観測モデルがガウス分布でよく近似できるような状況では、さほどたくさんの粒子は必要ない。一方、観測の非ガウス性が強い場合、アンサンブル変換カルマンフィルタによって得られる分布が本来の事後分布とはかけ離れたものになるため、importance sampling 時に実質的に推定に寄与する粒子の数が減少しやすくなり、これを防ぐために大量の粒子が必要となる。そこで、importance sampling 時に実質的に寄与している粒子の数の指標となる有効粒子数の値に着目する。図8の赤い線は、全粒子数に対する有効粒子数の比が、推定の初期段階でどのように変化するかを Lorenz 96 モデルによる実験で調べた結果を示す。また青の線は、推定値と真の値との差の標準偏差である。この実験では観測に非線型性を入れており、推定値と真の値との乖離が大きくなると事後分布のガウス分

布による近似が難しくなる。そのため、推定値が真の値から大きく外れている段階では、全粒子数に対して有効粒子数が少なくなる。しかし、推定がうまく行くようになると、全粒子数に対する有効粒子数は比較的大きな値で維持されるようになる。この例のように、事後分布のガウス分布による近似がうまくいかないときには、有効粒子数が全粒子数に対して小さくなり、結果として大量の粒子が必要となる。そこで、事後分布の計算において有効粒子数を参照することにし、十分な精度が確保できるまで粒子数を増やして有効粒子数が一定の水準に達したところで粒子の生成をやめるという方法で、importance sampling の際に用いる粒子の数を適応的に変化させるという方法を提案している。

(3) 2009 年のインフルエンザ感染動向調査を用いた外部地域からの影響の推定

はじめに

市民に対する警告を促すために、感染症研究所ではインフルエンザ警報・注意報を発表している。感染者数の予測に基づくが、予測のもとになるデータは全国 5,000 箇所の調査協力医療機関（定点）における週毎の新規感染者数である。我々の研究は、適切な感染伝播の力学モデルを導入することで、感染者数の予測を、特に学級閉鎖やワクチン接種などの介入を行ったときの予測を、改良することを目指している。今回の研究報告では、その準備として行った都道府県毎の動向（定点当たり感染者数）にどの程度外部からの影響が含まれているかを推定した結果を報告する。

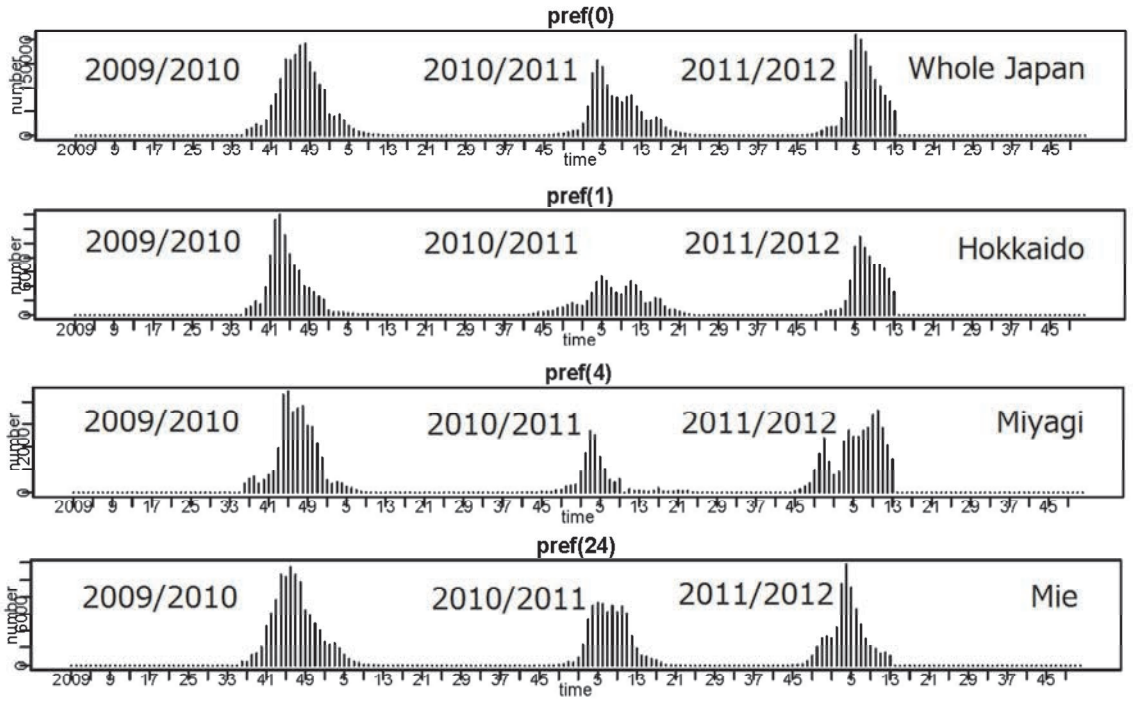

図1： 全国および 3 都道府県（北海道、宮城、三重)でのインフルエンザ様疾患定点当例数

定点観測データに見られる特徴

前述のように全国 5,000 箇所の定点で 1 週間ごとの新たに発生した感染者数が報告されているが、公開されているのはそれを都道府県別に集約した 48 の時系列である。特徴を概観するために図 1 に例を示す。県や全国レベルでは、感染パターンはいわゆるシングル・ピークである。特に 2009 年は、パンデミック・ウィルスが支配的であったためにこの傾向が顕著であったと考えられる。他のシーズンでは、複数のピークが見られるが感染伝播のダイナミクスを反映しているのか、複数の株種が異なる時期に流行しているのかを時系列だけから判断するのは難しい。そこで、解析上、理想的な感染者数変化をする

2009 年～2010 年シーズンを以下の解析では対象にする。概形ではシングル・ピーク型の感染パターンを持つものの、より詳細に観察すると急激な増加がみられる。これらの急激な増加は複数の県で同期している場合があることから、地域間の交互作用をモデルに取り入れる必要があると考えられる。今回の報告で、データから外部地域の影響が抽出できるかを調べようとした動機はここにある。

分析に用いるモデル

SIR モデルに確率的な変動を追加したものを用いる。SIR モデルは感受性人口（susceptible)、感染者（infected)、除外（removed; 免疫獲得のため感染伝播に寄与しない集団）の人口についての常微分方程式である。具体的には、

$$\frac{dS}{dt} = -\lambda IS, \qquad \frac{dI}{dt} = \lambda IS - \gamma I, \qquad \frac{dR}{dt} = \gamma I \tag{1}$$

という形式を持ち、λは感染力を表すパラメータ、1/γは平均罹患期間である。確率的変動は、式(1)を差分化したものに追加する。

$$\begin{aligned} S_n &= S_{n-1} - \lambda IS\Delta t \\ I_n &= I_{n-1} + \lambda IS\Delta t - \gamma I\Delta t + \Delta I_n, \end{aligned} \tag{2}$$

ここで、Δt は差分化の時間間隔、n は時点で、$t = n\Delta t + t_0$ が観測データが取得された時点と一致するように定義する。また、確率的変動 ΔI_n は、ある小さい確率 η で 0 でない値を取り、それ以外の場合は 0 を取るとする。確率分布として書くと

$$p(\Delta I) = \eta N(\Delta I \mid 0, \sigma_{\Delta I}^2) + (1-\eta)\delta(\Delta I - 0) \tag{3}$$

となる。解析の対象の時系列の概形はシングル・ピーク、すなわち無摂動の SIR モデルで説明されると仮定しているので、ΔI_n はほとんどの時点で 0 としている。

SIR モデルの解と実際のデータとの対応関係を見てみよう。SIR モデルによる新規感染者数の見積もり（これを J_n と書こう）は、感受性者から感染者になった人を積分することで得られる。すなわち、

$$J_n = \int_{t-\Delta t}^{t} \lambda SI dt \approx \lambda SI\Delta t, \quad t = n\Delta t + t_0 \tag{4}$$

他方、動向調査に記録される値（これを J_n^{obs} と書こう）は J_n に検出効率（感染者が医療機関に診察を受ける必要がある）を乗じたものと対応する。しかるに、検出効率をシーズンを通じて一定と仮定すると、適当なパラメータと初期条件のスケールによって Δt とともに省略できる。この仮定のもとで、J_n と J_n^{obs} とを対応づけるモデル（観測モデル）として

$$p(J_n^{\mathrm{obs}} \mid J_n) = N(J_n^{\mathrm{obs}} - J_n, \sigma_{\mathrm{obs}}^2) \tag{5}$$

を用いる。式(2),(3),(5)は状態空間モデルをなし、時系列全体の対数尤度は

$$\begin{aligned} \log p(y_N, \cdots, y_1 \mid \theta) &= \sum_{n=1}^{N} \log \int p(y_n \mid x_n) p(x_n \mid x_{n-1}, \theta) dx_n, \\ \theta = (\lambda, \gamma, x_0), \quad x_n &= (S_n, I_n, \Delta I_n, J_n), \quad y_n = J_n^{\mathrm{obs}} \end{aligned} \tag{6}$$

で与えられる。

解析手法

式(6)を θ について最大化して、その θ の下での（系列$(x_n)_n$に含まれる）系列$(\Delta t_n)_n$を推定し、適当な計算式で$(\Delta t_n)_n$を要約したものを当該地域における外部からの影響の指標とする。式(2),(3)は非線形なので、式(6)に含まれている積分は粒子フィルタによるモンテカルロ計算で求める。その際、適当な事前分

布 $p(\theta)$を設定し、$\theta_n = \theta_{n-1}$ という漸化式を式(2)に加えた拡大モデルに粒子フィルタを適用して(x_n, θ_n)を同時に推定することで、θ の決定を代替するという方法がしばしば取られる。しかし、初期条件とパラメータの間の非線形性のために適切な $p(\theta)$の設定は困難である。そこで、以下の簡便な方法を取ることにした。

手順１．率的変動を０として、θを決定する。

手順２．手順１で決めたθのもとで、粒子フィルタを適用し、$(\Delta t_n)_n$を推定する。

手順１を実行するにあたり、SIR モデルが解析解を持つことを利用して感染パターンの特徴と関連付ける。観測データを分析して、特徴パラメータ D_0, D_+, D_-,を抽出し、SIR モデルの解と対応付ける。

$$
\begin{aligned}
D_0 &\approx \max J(t) && \text{（ピーク値）} \\
D_- &\approx -\frac{d}{dt}\ln J(t)\Big|_{t\to-\infty} && \text{（開始点の周辺での傾き）} \\
D_+ &\approx -\frac{d}{dt}\ln J(t)\Big|_{t\to+\infty} && \text{（終了点の周辺での傾き）}
\end{aligned}
\tag{7}
$$

式(1)から t を消去して S を独立変数とする微分方程式を導き、これを積分すると

$$
I(t) - I(0) = S(0) - S(t) + \frac{\gamma}{\lambda}\ln\left(\frac{S(t)}{S(0)}\right) \tag{8}
$$

が得られる。この式を $t \to \pm\infty$で評価すると $I(t) \to 0$ より

$$
S_\pm = S_0 + I_0 + (S_0 - I_0)\ln[S_\pm / S_0] \tag{9}
$$

(S_0, I_0)を新規感染者数がピークとなる点にとり、近似 $J \approx \lambda SI$ を用いると、$dJ/dt = 0$ と式(7),(9)は 6 変数 $S_0, I_0, S_+, S_-, \gamma, \lambda$ についての 6 個の連立方程式をなす。簡単な計算によって、これらの方程式は 2 変数 S_0, γ の連立方程式

$$
D_\pm S_0 = 2D_0 + \gamma S_0 \log\left(\frac{S_0(D_\pm + \gamma)}{D_0 + \gamma S_0}\right) \tag{10}
$$

に還元される。この方程式をニュートン法などの数値解法で解くと、初期条件とパラメータをすべて決定することができる。しかし、式(7)の対応づけが有効なのはかなり理想的なデータ（シーズンの端が善く記録される）場合のみである。そのため、実際には式(10)の解の周辺でのランダムサンプリングにて尤度の式(5)を改良する。

確率的変動を要約した指標 ε をつぎのように定義する。当該地域とは別の地域を考え、対応する変数を$(\tilde{S},\tilde{I})$とする。このとき$\tilde{I}$の内の割合 $\varepsilon \in [0,1]$が、当該地域の感染伝達に影響し、それが ΔI_n として現れると仮定する。この仮定は、

$$
-\left(\frac{dS}{dt}\right)_{\tilde{I}} = \lambda S \cdot (\bullet\tilde{I}) = \sum_{i \in \mathrm{I}_{\mathrm{obs}}} \Delta I_i \cdot \delta(t - t_i) \tag{11}
$$

と表現できる。さらに、この別の地域が当該地域とほとんど同じように振る舞うと仮定して、$\tilde{I}$のかわりにIを使って、シーズンにわたる絶対値の平均値を ε とする。

$$\varepsilon = \frac{\sum_{i \in I_{\mathrm{obs}}} |\Delta I_i| / \gamma}{\int_{-\infty}^{\infty} I(t)dt} = \frac{\sum_{i \in I_{\mathrm{obs}}} |\Delta I_i|}{S(-\infty) - S(+\infty)} \qquad (12)$$

ε の直感的な意味は、累積感染者に占める確率的変動による感染者の割合である。

結　果

上述の手順で、各都道府県で J_n と ΔI_n を推定した。北海道と宮城県の例を図 2 に示す。上段が J_n（オレンジ線:無摂動解、赤線:確率的変動あり）および J_n^{obs} の（青棒）の時系列、下段が I_n（緑線）、および ΔI_n（緑棒、4 倍強調表示）である。確率的変動を加えたバージョンでは、観測データに良く追随し、図中右上に示された ε の値が大きくないこと（それぞれ約 0.25, 0.15）から確率的変動によるオーバーフィットにはなっていないことがわかる。いっぽう、図 3 に示す沖縄県の場合は、明らかにピークが複数あり SIR モデルでは説明できないため、観測データを再現するには確率的変動を多くする必要があり、ε = 0.52 という大きな値になっている。図 2 では、2 例だけを確認したが、各都道府県での ε の推定値は図 4 に示すように、ほとんどの県で 0.2 より小さく、やはりオーバーフィットになってないことが確認できる。

今回の研究報告では、地域別に独立な SIR モデルへの当てはめによって、他地域からの影響が 2 割程度であることを確認した。しかし、時系列の観察から複数県の間での連動した急激な変動の存在が定性的には確認されるものの、マルチコンパートメント SIR モデルのノード（地域）間結合定数を推定することは困難である。実際、疑似データを使って推定実験をすると他地域からの影響の総和は推定できても、その配分は推定できないことを確認している。今後は、地域間流動表などの疫学以外のデータを使って、感染ネットワーク・モデルの構成を進めたい。

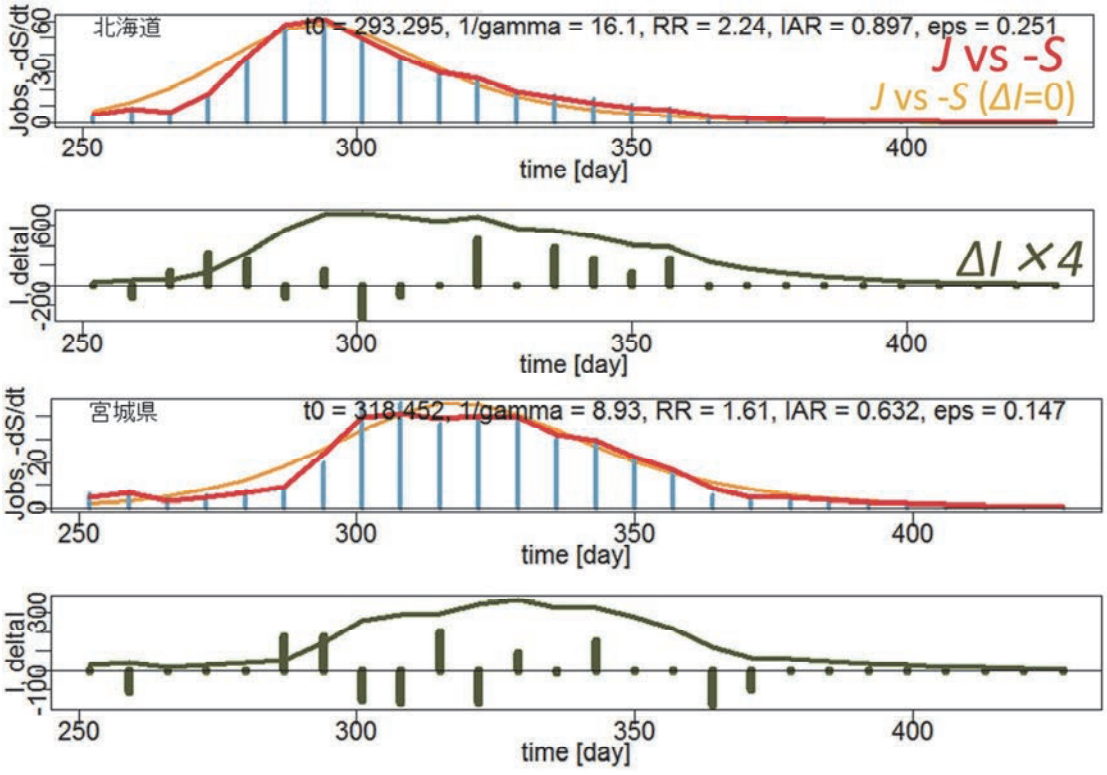

図2： J_n と ΔI_n の推定事例（その 1）