ビッグデータ価値化への挑戦
薬剤副作用分析と航空機着陸システムの安全性設計から

森永　聡（日本電気株式会社）　青木　健児（日本電気株式会社）
鈴木　和史（日本電気株式会社）　藤巻　遼平（NEC Laboratories）
福島　荘之介（電子航法研究所）

概要　本論文ではビッグデータ価値化への挑戦の活動について，薬剤副作用分析と航空機着陸システムの安全性設計を例にとり，プラクティカルな面に重きをおいて報告する．前者は診療報酬明細のデータから薬剤の副作用発現を同定する試み，後者はGPSを利用した航空機着陸システムでGPS信号の信頼範囲を分析する方式の開発である．どちらも，タスクの適用領域の専門家とデータ分析の専門家の密接な連携により，初めて意味ある進捗を得られたものである．

1. はじめに

本論文では，筆者らが業務として取り組んでいる，ビッグデータ価値化への挑戦に関する活動について，副作用マイニングと着陸誘導システムの異常検出という具体的な二つのタスクを題材に，プラクティカルな面に重きをおいて紹介する．

近年，大量のデータから価値ある情報を抽出し，実社会で活用する試みが極めて盛んになり，ビッグデータというキーワードとともに一つの潮流となっている．例えばEコマースの分野では，大量の購買履歴データからユーザの購買行動に関する傾向等を抽出し，商品推薦に活用することは既に広く実用化されているし，インターネット上の大量の書き込みを分析することにより，マーケティング関連の知見を抽出するサービスも商用化されている．一方，大量のヘルスケアデータを分析することによる医療や保健に関する知見抽出の試みや，機械や工場等の挙動をセンサによって大量収集したデータを分析することによる自動監視や制御の試みも広く実施されている（例えば，参考文献1）や2））．

これらの試みの多くにおいては，機械学習に基づくデータマイニングのアプローチが採用されている．そこでは大まかに1)各種分析処理がうまくいくように生データを整形・変換する，2）変換後のデータ中で成立している法則や傾向・パターンを機械学習技術により自動抽出する，3）抽出した法則やパターンを利用して，数値予測やカテゴリ判別，異常検出や制御といった顧客価値を実現する，といった分析プロセスが踏まれる．1）で生成される2）の入力データは属性データと呼ばれる．

機械学習に基づくアプローチによるビッグデータ価値化の取り組みとは，具体的な案件に対して上記1)，2)，3）の分析プロセスを適切に設計することに尽きる．その成功のためには，単なる分析技術的な側面にとどまらず，適用分野の背景知識を踏まえた適切な問題構成や，現実的なコストの下でのデータ準備や評価といった実務的な側面も非常に重要となる．以下，第2章では薬剤の副作用分析，第3章では航空機着陸システムの安全性設計といった全く異なる分野のタスクを題材に，プラクティカルな視点から筆者らの経験を記述する．また最後に第4章で，これらに共通する課題の整理に挑戦する．

2. レセプトからの薬剤副作用分析

2.1 レセプトデータ

レセプトとは診療報酬明細書のことで，医療機関から健康保険組合等への医療費の請求書である．基本的に，各医療機関より患者毎に一カ月単位で発行されており，患者名や性別，病院名や診療科，傷病名，診療行為名，薬剤名，請求金額（点数）などが記載されている．薬剤や診療行為ごとに診療報酬点数が決められており，医療機関はこの点数を合算して保険者に医療費を請求する．

日本では健康保険に関しては制度上，国民皆保険であるので，1億2千万人以上の，それもほぼ全ての医療事象に関して記録されたデータが随時発生していることになり，まさにビッグデータとよぶにふさわしい．特に，厚生労働省保険局が構築しているナショナル・レセプト・データベース[3]は，各医療機関が個別に管理している診療記録やレセプト情報を，電子化したデータベースとして国が一元的に管理・運用するもので，医療費適正化計画の作成等に資する調査・分析や，医療サービスの質の向上等を目指した正確なエビデンスに基づく施策の

推進への利用等についても，厚生労働省においてデータ利用の公益性等に関し検討が進められている．

本章においては，レセプトデータの価値化への挑戦の一例として，独立行政法人医薬品医療機器総合機構（以下 PMDA と略記）からの委託業務として行った，レセプトデータからの薬剤副作用事象の検出について紹介する．本章の学術面・技術面の内容は参考文献 4）からの引用であり，それらに関する詳細は当該文献を参照されたい．また，本章の内容は PMDA の見解を必ずしも表すものではなく，筆者らのそれである．

2.2 薬剤副作用情報の抽出

PMDA は，医薬品等の安全対策業務の強化・充実策の一環として，電子診療情報等を安全対策へ活用する事業を MIHARI Project として立ち上げており，その活動の一つとして，レセプトデータからの薬剤副作用事象の検出を検討している[5]．現状，市販後の医薬品に対する副作用事象に関する監視が，医療現場や一般使用者等からの自発的な報告の分析を中心に行われているのに対し，大量に発生するレセプトデータを分析することにより，それを補完することが本章の目的である．特に，未知の副作用事象の発現に関して幅広く迅速に検出することが究極の目標となる[4]．

このような目的を達成するためのデータ価値化への挑戦アプローチとして，筆者らは大きく分けて検討フェーズを二つに分解した．検討フェーズの第一段階は，レセプトデータが，副作用発現に関する情報を本当に持っているかの確認である．レセプトには副作用発現の有無は明示的には記載されていないので，副作用に関する情報を得るためには，副作用発現の有無でデータ傾向等に何らかの差が生じている必要がある．その検証のために，まず，副作用事象が発現している患者とその時期を，レセプトデータから同定可能かを確認するための分析を行った．検討フェーズの第二段階は，レセプトデータから副作用情報を抽出する方式の開発である．PMDA の目的を達成するためには，「どの薬が，どのような副作用を起こしているか」に関する情報を抽出する必要がある．さらに当該副作用の治療等に用いられた薬剤や診療行為に関する情報も同時提示することが望まれるため，副作用仮説として（原因薬，治療薬，診療行為，傷病名）の四つ組を自動抽出する方式の検討を行っている．

上記のとおり，両フェーズとも技術的には極めてシンプルなことをやっているともいえるが，実際に検討を実施しようとすると実務的には極めて大きな困難が伴った．

第一の困難は，検討の手掛かりや結果評価に不可欠な正解データの入手である．すなわち，特定の患者の特定の時期において実際に副作用事象が発現していたかどうかや，副作用仮説として抽出された特定の四つ組が本当に副作用であるかどうかを確認するのは，高度に医学・薬学的な知見に基づいた人手による判断に基本的に帰着されてしまう．

第二の困難は，各患者における医療事象の羅列であるレセプトデータから，どのように属性データを作ればよいのかも全く自明でないことである．検討フェーズ 1 では，副作用発現の有無で差異が出そうな何らかのレセプトデータの特徴を定式化して属性データとする必要があるが，そのような特徴の候補を挙げるためには，医学・薬学の知見に加え，属性データを入力する機械学習技術に関する知見も必要になる．結局これらの困難な点は，医学・薬学といった適用分野の領域専門家による粘り強い作業や，分析技術の専門家との互いに相手の分野を理解する努力を伴いながらの密接な協働によって乗り越えるしかなかった．

さらに，第三の困難として，実際のレセプトデータは，いわゆる表記ゆれ（同一の医療事象でも，レセプト作成者によって表現が違い，データ上では別の値として現れる）の問題も極めて大きい．しかしながら今回の検討では，その問題の解決をスコープ外とするために，株式会社日本医療データセンターにより必要な整理・修正や個人情報の匿名化が施されたデータセット（約 40 万人分，データ期間は 2005 年 1 月～2008 年 12 月）を，PMDA から貸与をうけて用いた．このようなデータ整備作業も，極めて高度な医学的・薬学的知見とデータ分析の技術的側面の両方を見据えて行われるべきものであり，領域専門家と分析専門家が密接に協力して進める必要がある．

以下では検討フェーズ 1 の発現同定について詳しく説明する．検討フェーズ 2 の副作用仮説抽出は現在進行中であり，今後の課題として簡単に紹介する．

2.3 副作用事象発現の同定

検討フェーズ 1 においては，副作用事象が発現している患者とその時期を，レセプトデータから同定可能かを検討する．具体的には，指定された患者 ID と時期のペアに対して，レセプトデータを参照したうえで「副作用事象が発現しているかどうか」を出力するシグナル検出器を構成し，その精度を評価するという手順を踏むこととした．ここで，時期としてはレセプトの発行単位に合わせ「年月」という時間分解能で指定することとしタイムスロットと呼ぶことにする．シグナルは YES/NO の 2 値ではなく，発現の疑わしさを表す 0 から 1 の連続値と

する．すなわち，入力として患者 ID とタイムスロットのペア X=(PID,yyyymm)を与えると，副作用発生確率 Y∈[0,1]を出力するシグナル検出器を構成して評価することとなる．

もし，あらゆる副作用に関する原因薬と症状等の組合せに関する網羅的な情報が入手できるのであれば，それに該当するかの判断をベースにすることで，シグナル検出器は比較的簡単に構成できるであろうが，そのような情報は存在していない．特に，本件の目標は未知の副作用に関しても検出することであるので，入力 X に相当するレセプトデータが，既に知られている副作用の原因薬や症状に該当するかのマッチングをベースにしたシグナル検出器は想定外として検討を進める．

まず，正解データの作成や属性の作り方検討の効率向上のために，代表的な副作用に関する（原因薬，治療薬，診療行為，副作用傷病）の四つ組の情報のリストアップと，単なる医療事象のログであるレセプトデータを見やすくするための可視化を行った．

四つ組情報のリストアップは，痙攣や劇症肝炎・肝不全，間質性肺炎や脳梗塞といった副作用のタイプ別に，原因となる薬効群，治療薬，診療行為，副作用傷病名のそれぞれの項目をまとめたもので，PMDA が臨床的知見をもとに作成したものである．以降，このリストアップされたデータを副作用辞書とよぶ．図 1 は副作用タイプの中で痙攣に対する副作用辞書の例であり，表の該当要素からなる四つ組すべての組合せ（17×3×27×6 通り）が，このタイプの副作用を表すものである．逆に，未知の副作用とは，副作用辞書に相当するものがない組合せのものであるとする．

図 2 はレセプトデータを可視化した例である．ある一人の患者について，時系列方向に各種の医療事象の発生の有無を+-の記号パターンで表現している．特に副作用辞書に該当する医療事象に関しては有無を表す記号を@-に変えて上部にまとめて再掲する（下部に通常の+-で表現されたものも表示されている）ことにし，直観的な概観性をあげている（詳細は参考文献 4）参照）．

このような準備等をもとに，PMDA の専門的知見に基づいて，患者 ID とタイムスロットの組合せに対するシグナルの正解データが付与された．別の観点から複数の正解セットを作成したが，主に副作用辞書に該当する医療事象の発生パターンに着目して作成された，正例（副作用が発現しているとみなせる患者 ID とタイムスロットの組合せ）219 件，負例（そうではない組合せ）29782 件を用いた実験をここでは例にとって紹介する．

シグナル検出器の属性データとしては，領域専門家である PMDA と分析専門家である筆者らでレセプトの可視化結果をもとに議論を繰り返した結果，指定された患者の指定されたタイムスロット周辺の時期における，医療事象の発生パターン，すなわち可視化した場合の+-記号のパターンをベースにすることとした．これは，副作用発現が起きるタイミングにおいては，当該患者において「それまで使用されていた薬剤が中止」「新たな傷病が発生」「なんらかの薬剤や診療行為が新規追加」がおきるだろうという知見に基づくものである．具体的には，当該患者の当該タイムスロット前後 6 カ月間を切り出し，+-記号の全パターン（64 通り）について該当する医薬品，

図 1　副作用辞書の例

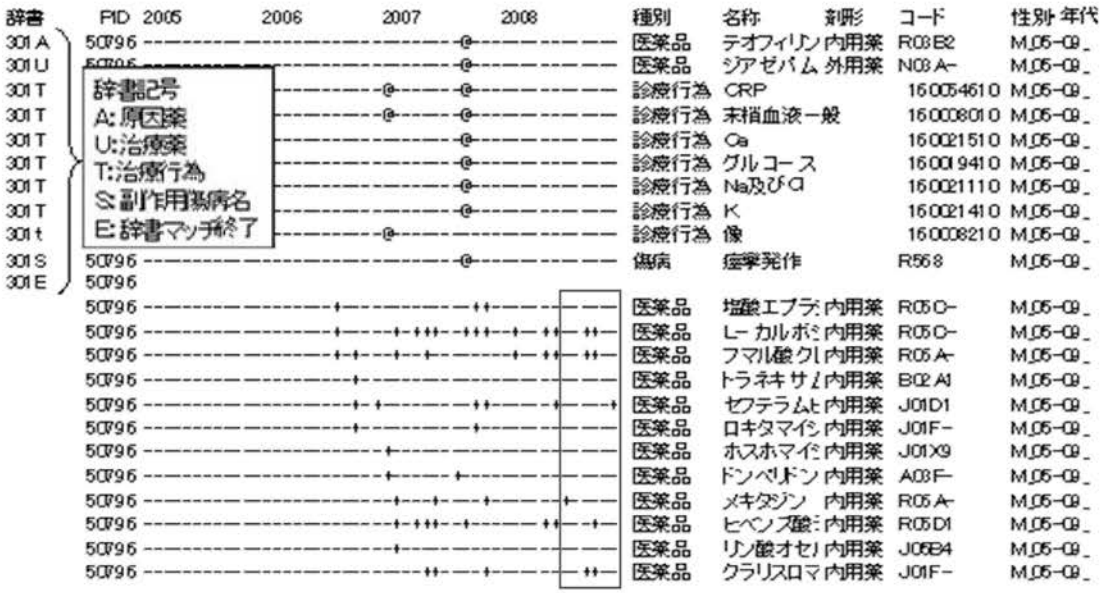

図 2　レセプトデータの可視化の例

診療行為，傷病の種類を集計したものを属性とした．前述の通り，既知の副作用に該当しているかの情報を使用するのは不適切であるので，上記の集計は可視化した場合の+-記号パターンのみに着目して行うことに相当する．例えば，図 2 のデータにおいては，図中四角内（2008 年下期頃）の 6 カ月間において，パターンが“------ ”となる医薬品が 6 種類，“--++--”が 3 種類，“-----+”が 1 種類，“+-----”が 1 種類，“---+--”が 1 種類となっているので, このタイムスロットの医薬品に関する 64 次元の属性ベクトルは，それぞれ上記に対応する成分が 6,3,1,1,1，残りは 0 となるように構成する．

正解データ（正例と負例の属性データ）を学習用セットと評価用セットに分割し，前者を用いてシグナル検出器の各パラメータを学習によって調整し，後者を用いてその精度評価を行った．シグナル検出器としては，検討フェーズ 2 での知見の再利用などを見据えて，解釈性の高い L1 正則化ロジスティック回帰モデルを採用した．

図 3 はシグナルの精度を評価セットで評価した結果である．検出器が出力するシグナルはアナログ値であり，副作用発現のアラームとみなす閾値を変えたときの，正例的中率と網羅率の関係を図 3 は示している．横軸は，評価用の正例の全件数に対する，スコア上位ランキング N 番目までを閾値とした場合に見つかった比率であり，縦軸は，スコア上位ランキング N 位 までの中に見つかった正解の数なので，なるべく左側に山が高くなっているのが望ましいグラフである．一般的に，閾値を高く設定するほど的中率は上がるが網羅率は下がる．シグナル強度の上位 100 件をアラームとみなすと 14 件が正例として的中し，300 件であれば 22 件となる．この結果においては，正解セットにおける正例の比率が 1%未満であるにも関わらず, 図の左側で多くの正解を的中させている．これはレセプトデータは副作用発現に関する情報を含んでおり，副作用発現の有無でデータの傾向に差があることを示している．筆者らは，検討フェーズ 1 は一定のレベルで成功したとし，フェーズ 2 の副作用仮説抽出へと進んだ．

2.4 副作用仮説の抽出

検討フェーズ 1 において，副作用発現の有無に応じてレセプトデータに傾向の差があることが確認できたが，副作用発現アラームを一定の精度で発出するシグナル検出器は，その傾向の差を正解データから学習することに成功しているといえる．そこで，フェーズ 2 では，シグナル検出器の出すアラームと，シグナル検出器が学習した傾向に着目して，副作用仮説を自動抽出する方式を検討したが，思わしい結果は得られていない（詳細な手順や評価結果，考察等は参考文献 4）を参照）．問題の定式化や，それに適した属性の定義，必要な情報の整備の各面において，領域・分析の両専門家が協力して新たな手法の開発を進めていく必要がある．

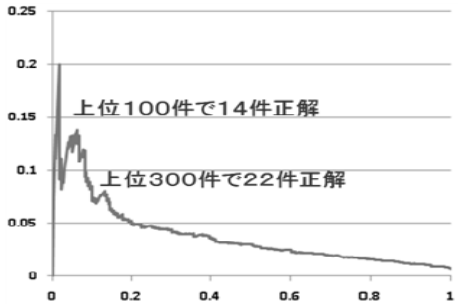

図 3　シグナル検出器の精度

縦軸：正例的中率　横軸：正例網羅率

3. 航空機着陸システムの安全性設計

3.1 次世代航空機着陸システム GBAS

GBAS（Ground Based Augmentation System: 地上型衛星航法補強システム）は，GPS を利用した次世代の航空機着陸システムである．2000 年代前半に世界標準が制定され，米国，ドイツを筆頭に世界各国において導入に向けた活動が行われてきた．そして 2012 年 2 月にドイツ・ブレーメン空港にて世界初となる GBAS 運用が開始された．さらに 2012 年末から 2013 年初頭にかけて米国の 2 空港，スペインの 1 空港で運用開始が予定されており，システムの実証段階から導入段階へと移行が進んでいるところである．日本においても，電子航法研究所により 1990 年代後半から研究が行われており，国内初となるプロトタイプが 2010 年に関西国際空港に設置された（参考文献 6, 7）．Boeing 社の最新鋭機として話題の B787 型機は GBAS 機上装置を標準装備している．この B787 型機を使用して，2011 年 10 月に電子航法研究所と全日空とが共同で，2012 年 4 月に同研究所と日本航空とが共同で，関西空港にて GBAS フライト試験を成功裏に終えているなど，GBAS 実現への技術的目処がつきつつある[8].

GBAS の構成，および動作原理について述べる．図 4 は GBAS のイメージ図である. GBAS は GPS 等の測位衛星，地上システム，機上システムから構成される（それぞれ図中①，②，③）．測位衛星の性能は自動車や歩行者が自位置を知るためには十分であるが，航空機の着陸支援のためには不十分である．このため地上システムが必要となる．地上システムには大きく分けて以下の三つの機能がある：1）GPS 信号に含まれる誤差を推定し，その推定値を航空機に対して提供する機能，2）上記推定値の信頼範囲を提供する機能，3）航空機に対して最終進入経路を提供する機能．高い安全性が要求される GBAS において最も重要なのが 2）であり，3.2 節で詳述する．機上装置は，これらの情報をもとに最終進入経路からの差分をパイロットに提供する．また，地上システムから提供される推定値の信頼範囲を使用して，測位結果の信頼範囲を計算する．計算した信頼範囲が予め定められている許容値を超えている場合は，十分な安全性が得られないため，GBAS 以外の手段に切替えるか，または進入着陸そのものを回避し進入復行を行う．

GBAS のメリットについて述べる．従来の航空機着陸システムでは地上システムからの誘導電波により航空機を着陸させるため，以下の制限があった：1）大規模な施設およびその設置のための用地の確保，2）地上設置型のため装置一式につき一つの進入経路かつ滑走路に対して直線進入のみサービス提供可能．一方，GBAS では測位衛星によって自位置の計算を行い，地上システムからのデジタル通信により進入経路を指示するため，上記の制限を受けない．これにより，より短い進入経路，住宅地や山岳地を回避した進入経路の実現（その結果として燃料消費・飛行時間の削減，騒音低減等のメリット）が期待される．また，地上装置の所要用地の削減および装置一式で複数の進入経路にサービス提供可能というメリットも存在する．

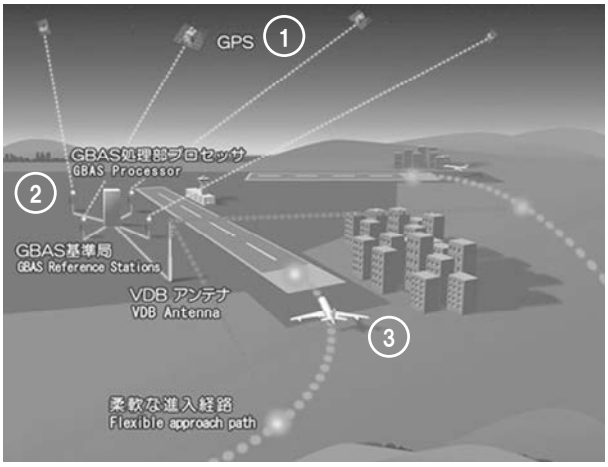

図 4　GBAS の構成

3.2 GBAS におけるデータ分析の課題

GBAS 地上システムのうち最も重要な機能である GPS 信号誤差の推定値の信頼範囲を計算する機能について説明する．GPS による測位誤差は，衛星と受信機間の距離情報や衛星位置情報の誤差が原因となって現れる．距離情報の誤差の要因としては，GPS からの電波が電離圏や対流圏を通過するときに生じる伝搬遅延，GPS 衛星の内蔵時計がずれていることによる衛星時計誤差，GPS 受信機周辺の地上反射波によるマルチパス等がある．衛星位置情報の誤差は GPS 自身が放送している軌道情報に内在する誤差である．GBAS 地上システムは，これらの誤差の合計の推定値，および各誤差の推定値に関する信頼範囲を機上システムへ提供する．機上システム側は誤差の推定値を使用して高精度に自位置を求める．同時に，誤差の推定値の信頼範囲から自身の位置の信頼範囲を計算し，その信頼範囲が空港や進入経路ごとに予め定められている許容値を超えていないかを判断する（安全性の確認）．なお，本稿では詳細は省略するが，地上システム側のもう一つの重要な機能として，推定した誤差が信頼範囲内に収まっているか（突発的な異常事象が発生していないか）の監視をリアルタイムで行うインテグリテ

ィモニタと呼ばれる機能がある．

次に，GBAS におけるもう一つの重要な性能要件であり，安全性とトレードオフの関係にある可用性について簡単に述べる．先に述べた安全性のみを満足すればよいならば，信頼範囲を大きめに見積もっておけばよい．信頼範囲を大きく見積もれば見積もるほど，実際の誤差が信頼範囲を超える可能性が低くなるからである．しかしその一方で，信頼範囲を大きくすることは信頼範囲が許容値を超える可能性が高くなることを意味し，信頼範囲の過度な増大はGBASサービスの可用性の無用な低下を招く．

GBAS では，このようにトレードオフの関係にある安全性，および可用性のそれぞれについて数値的に最低要求値が与えられている[9]．この要求値を表 1 に示す．安全性要件は，「測位誤差が信頼範囲を超えない確率」で与えられている．測位誤差が信頼範囲を超えない確率が“$1-2\times10^{-7}$ in any approach” とは，考え得るいかなる厳しい条件下での着陸においても，着陸中に測位誤差が信頼範囲を超える確率は 2×10^{-7} 以下ということであり，非常に厳しい要求であることがわかる．前述のように，測位誤差の信頼範囲は地上システムが提供する GPS 信号誤差の信頼範囲をもとに機上システムが計算する．地上システムには，機上システムの測位誤差が，その信頼範囲を超えることがないように GPS 信号誤差の保守的な信頼範囲を提供することが求められる．可用性要件は「GBAS による着陸が可能な時間帯の割合」が “99%～99.999%” 以上であることが求められている．要求値に範囲があるのは，GBAS が設置された空港に着陸する航空機の数，または代替着陸手段の有無，といった運用環境により求められる要件が異なるためである．地上システムには，この安全性要件と可用性要件とを同時に満足するために，実際の誤差を確実にカバーしつつ，かつできるだけ小さな信頼範囲を提供することが求められる．

表 1　GBAS 性能要件

項目	性能要求値
安全性（Integrity）	測位誤差が信頼範囲を超えない確率： $1-2\times10^{-7}$ in any Approach
可用性（Availability）	GBAS による着陸が可能な時間帯の割合： 99%～99.999% （運用環境による）

3.3 マルチパスによる信号誤差

GBAS 地上システムは，前節に述べた信号誤差の推定値の信頼範囲を 3 つの成分に分けて提供している：1）地上システムに起因する成分，2）電離圏に関連する成分，3）対流圏に対応する成分．本節では，このうち地上システムに起因する成分について詳しく述べる．

地上システムに起因する成分の大部分がマルチパスによる誤差である．マルチパス誤差とは，地上システムに備わっている GPS アンテナが，衛星からの信号（直接波）のみでなく，地面や建物等の複数の経路からの信号（反射波）を同時に受信していることにより生じる，GPS 信号の測定誤差である．地上システムに起因する成分には，マルチパスのほか，受信機雑音誤差，アンテナ特性による誤差等がある．以降では，表現の簡便さの観点から，これらを含めてマルチパス誤差と呼ぶ．地上システムでのマルチパス誤差によって，機上システムへ提供する誤差推定値に誤差が生じ（誤差の推定に対してまた誤差が生まれてしまうことになり），機上システムの測位精度の劣化につながる．従って，機上システム側で自身の位置の信頼範囲を計算する際に，地上システムのマルチパス誤差の影響を適切に加味する必要がある．そのために地上システムにはマルチパス誤差の大きさを適切にモデル化し，誤差の保守的な信頼範囲を機上システムに対して提供することが求められている．

図 5 に国内のある空港で実際に取得したマルチパス誤差の推定値と地上から見た衛星の仰角および方位角の関係のデータを示す．これらのデータは以下の手順によって得られる．地上システム側に 100m 程度の間隔で設置された 3 機または 4 機の受信機によって，GPS 信号の観測を行う．そして，各受信機で観測された GPS 信号に含まれる誤差を推定し，これらの推定値から複数受信機間で共通の誤差成分を取り除く．これにより，各受信機に対して固有に影響を与えるマルチパス誤差を推定することが可能となる．受信機間を 100m 程度離して設置するのは，複数受信機に共通となるマルチパスが入らないようにするためである．また，ここで示されているデータは平均が 0 になるように予め基準化されている．図 5 から，マルチパス誤差の大きさは±20cm 程度を見積もっておけば十分でありそうなことが分かる．この程度の誤差は航空機の安全な着陸に影響を与えるものではない．GBAS の安全性保証で重要なのは，マルチパス誤差としてこの程度の大きさの誤差が内在することを考慮しておくこと，およびこの考慮から外れる大きさの誤差は何らかの手段により検出し排除することである．

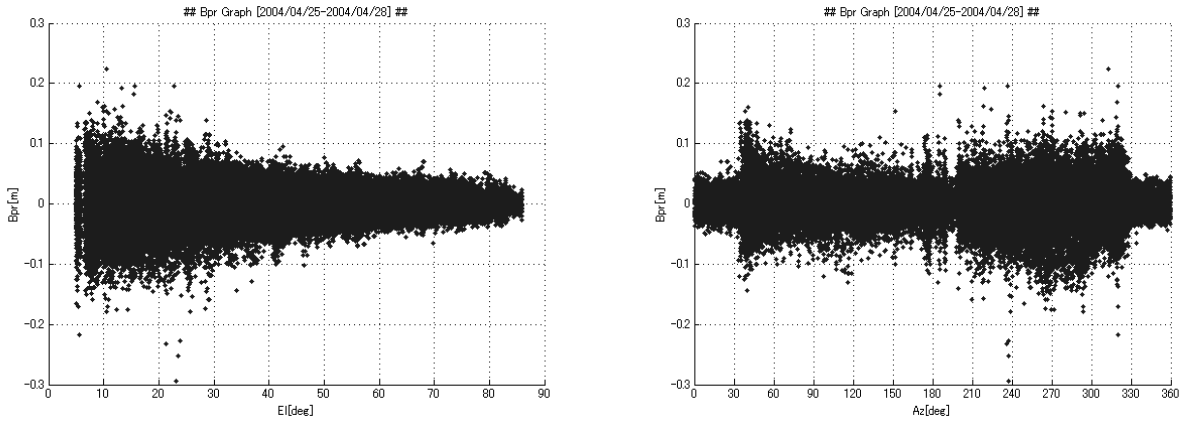

図 5　マルチパス誤差と方角の関係（左：仰角，右：方位角）縦軸：誤差の推定値，横軸：角度

図 5 を見るとマルチパス誤差の推定値の散らばりの大きさは仰角にも方位角にも依存していることが分かる．仰角に関しては高仰角ほど（90 度に近いほど）散らばりが小さくなる傾向が見て取れるが，単調に減少しているわけではない．また，減少の仕方は低仰角ほど（0 度に近いほど）急である．一方，方位角に関しては仰角のような規則性はなく，角度によって散らばりが非連続的に大きくなったり小さくなったりしている様子が見て取れる．前者に関しては，地上受信機から見た衛星の仰角が小さいほどマルチパスの影響が強くなり，また推定結果も不安定になることが一般に知られている．後者に関しては，空港周囲の地形や建物の影響により，複雑な依存性を示すものと考えられる．

先述したように，信頼範囲の過大／過小評価は GBAS サービスの可用性／安全性の低下に繋がるため，マルチパス誤差の推定値に対するより精密な信頼範囲の計算が求められる．本論文では，推定値の信頼範囲の計算を確率分布によるモデル化をもとに行うアプローチを考える．上記の観察より，マルチパス誤差の推定値の分散（散らばり）の仰角および方位角に対する複雑な依存性を確率モデルによって巧く表現することが精密な信頼範囲の計算のためのキーポイントとなる．

3.4 マルチパス誤差の高度なモデル化

従来の方法では，以下のような方法を用いてマルチパス誤差の推定値の分散を求めていた（図 6 にイメージ図を示す）：従来手法 1）仰角・方位角を予め決めておいた閾値によっていくつかの区間に区切り，その結果得られる矩形領域ごとに分散を求める，従来手法 2）仰角・方位角に関する分散の依存性を少数個のパラメータによって定まる曲線モデルによって表し，データからパラメータの値を求める．しかし，これらの手法には以下のような問題点があった：従来手法 1）分散の変化点に相当する閾値に関して予め固定された値を用いるため，空港ごとに特有なマルチパスの影響を表現することができない，従来手法 2）少数個のパラメータによって表される曲線では，実データに対して観測される分散の不規則な変化を表現することができない．上記の既存手法の問題はいずれもデータの構造に対する仮定が強すぎることに起因すると考えられる．一般にマルチパス誤差の影響は空港ごとに異なり，またその影響は特定の形に限定されたモデルによって表すには複雑すぎる．この問題を克服しない限り，新たな空港に GBAS を設置するたびごとにマルチパス誤差のモデル化のために相当の人的・時間的

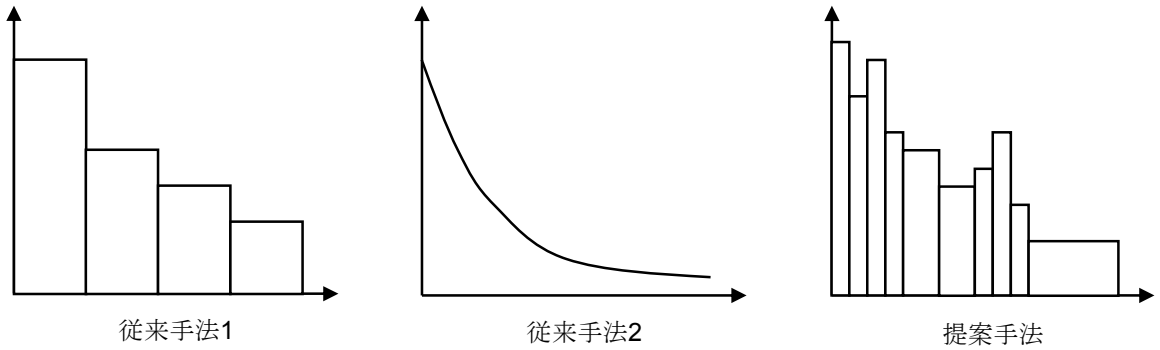

図 6　マルチパス誤差のモデル化の方法（縦軸：分散の大きさ，横軸：仰角・方位角）

コストが必要となる．

そこで筆者らは，これらの問題点を克服するため，ノンパラメトリックモデルと呼ばれる種類のモデル化手法を用いることにした．ノンパラメトリックモデルを適用すれば，与えられたデータによってモデルの形（ここでは分散の変化点を表す閾値の数と値）を柔軟に変化させることができる(図 6 の提案手法にイメージ図を示す)．詳細は省略するが，本研究では最適なモデルの形を決定するための基準として情報量規準と呼ばれる指標を用いることにした（情報量規準については、例えば参考文献 10）を参照）．

しかし，ここで問題となるのは，モデルの形の候補が指数オーダの組合せだけ存在することである.そのため，最適なモデルをナイーブに求めようとすると現実時間での計算は不可能となる．そこで筆者らは，動的計画法を用いた探索アルゴリズムを構成することにより，多項式オーダの時間で最適なモデルを見つけることを考えた．しかしこの方法では，仰角・方位角の閾値を同時に最適化することはできない．そこで，領域専門家と分析専門家の双方で議論を重ね，仰角・方位角のうち他方に対する依存性がより小さい仰角に対してまずは閾値の最適化を行い，結果として得られた仰角の区間ごとに方位角に対する閾値の最適化を行う，というヒューリスティックな戦略を採用することにした．

図 7 に提案手法を用いて実際に図 5 のデータの標準偏差（分散の正の平方根）をモデル化した結果を示す．低仰角から高仰角になるにつれ分散が小さくかつ閾値の間隔が広く設定されていることが見て取れる．また，分散は単調に減少するわけではなく低仰角側よりも大きな値となっている角度も存在する．方位角に関しては，低仰角から中仰角にかけて 150 度付近の分散が他の方位角に対する分散と比較して非連続的に小さくモデル化されていることが見て取れる．これらの結果が観測データの構造を正確にモデル化したものであるかどうかの定量的な評価は別途必要であるが，少なくとも既存手法ではこのような複雑な構造を推定結果として得ることはできない.

3.5 今後の課題

本章では，GBAS の安全性と可用性のトレードオフの観点から，マルチパス誤差の信頼範囲の精密な計算という GBAS におけるデータ分析の課題について述べた．本研究で提案した手法により，取得したマルチパス誤差データの構造を反映した最適な閾値を自動的に構築することが可能となった．これは，システム運用開始前のパラメータ調整作業や，積雪，草木の成長等によりマルチパス特性が変化した際のパラメータ更新作業の労力を低減することに繋がる．しかし，機械学習技術において全てのパラメータが自動的に最適化されるということは有り得ず，例えば，パラメータ更新の頻度・タイミングの決定などに関しては実運用下での様々なコストの観点から，領域・分析両専門家による議論の余地が残されている．

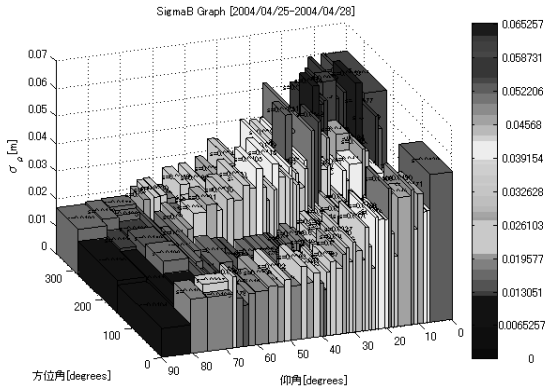

図 7　提案手法による標準偏差の推定結果

GBAS における他のミッションの一つとして，衛星故障，電離圏擾乱，一時的なマルチパスの増大などによる突発的な異常事象の監視を，衛星からの時系列データに対してリアルタイムで行う，というものがある．これは GBAS の安全性と継続性を同時に満たすことが要求され，GBAS におけるもう一つのトレードオフに関する課題である．筆者らはこの課題にも取り組んできたが，その過程で以下のような困難があった．第一に確率モデルの精度に対する要求レベルが極めて高いことがある．リアルタイムでの監視という本課題の性質上，モデルの精度が低いと頻繁に誤警報が上がることになり，継続性が簡単に破綻する（ここで，安全性の要求を満たすこともマルチパス誤差同様にやはり必須要件である）．この困難さの一因として，表 1 の安全性の項目で示されているように，極めて 1 に近い確率に対応する信頼範囲を求めなければならないことがある．この種の問題は統計学の分野で極値理論と呼ばれており，様々な応用領域で研究が行われている[11]．また，他の原因として，本活動事例のように様々な機能から構成されるシステムに対してデータ分析技術を組み込むようなタスクでは，モデルの精度の低さの原因を特定することが難しい，ということがあげられる．純粋にモデリングの問題なのか，データの前処理が不適切だったのか，または数値計算やバグといったプログラム上の問題なのかを調査する作業に多大な労力を要する．この観点からも各領域の専門家による協働が課題解決における重要な要素となる．

4. おわりに

レセプト副作用分析，GBAS 安全性設計のどちらにおいても，適用対象の領域専門家と分析技術の専門家の密接な協働により，はじめて意味のある進捗が実現できている．これに限らず，あらゆる実案件において両者の協働の成否がタスクの成否のカギとなり，さらに異文化交流の常として難しさのポイントにもなる．

協働が必要とされる場面は，完全性の高いデータ収集の基盤から，実社会における価値まで到達するための適切な問題の構成，正解データ等の必要な情報の作成，実世界からの要求に基づく制約の反映，分析結果の評価といった，タスクの実現に必要な活動のほぼ全てにわたる．筆者らの経験では，成功事例は，領域と分析の両専門家が互いに相手のテリトリーまで踏み込んで活動した場合に限られることを記して本報告の結びとしたい．

参考文献

1) H. C. Koh and G. Tan: ” Data Mining Applications in Healthcare”, Journal of Healthcare Information Management, Vol. 19, No. 2, PP. 64-72, (2011)
2) V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey”, ACM Computing Surveys, Vol. 41, No. 3, Article 15, (2009)
3) 厚生労働省: レセプト情報・特定健診等情報データベースの第三者提供について
http://www.kantei.go.jp/jp/singi/it2/iryoujyouhou/dai7/siryou2_1.pdf
4) 医薬品医療機器総合機構: 診療報酬明細書のデータを用いた，データマイニングに関する技術的検討業務（2010 年度検討報告書）
http://www.info.pmda.go.jp/kyoten_iyaku/file/e_rece-report1106_002.pdf
5) 医薬品医療機器総合機構: Mihari Project について
http://www.info.pmda.go.jp/kyoten_iyaku/mihari.html
6) 福島荘之介，齊藤真二，吉原貴之，齋藤享，藤田征吾: 地上型衛星航法補強システム（GBAS）の開発と安全性要求の保証，平成 24 年度 電子航法研究所研究発表会 講演概要，
http://www.enri.go.jp/report/hapichi/pdf2012/H24_07.pdf (2011 年 10 月 22 日現在)
7) 福島荘之介，工藤正博，齊藤真二，吉原貴之，齋藤享，藤田征吾，藤井直樹: 衛星航法による精密進入着陸システムの開発と安全性の保証，信学論 B, Vol. J94-B, No.7, pp.802-811, 2011 年 7 月，
http://search.ieice.org/bin/pdf_link.php?category=B&lang=J&year=2011&fname=j94-b_7_802&abst=. (2011 年 10 月 22 日現在).
8) 伊藤正宏，福島荘之介，山康博，齊藤真二，藤田征吾，長井丈宣，赤木宣道: B787 による GBAS プロトタイプ飛行実験，平成 24 年度 電子航法研究所研究発表会 講演概要，
http://www.enri.go.jp/report/hapichi/pdf2012/H24_08.pdf (2011 年 10 月 22 日現在)
9) ICAO: International Standards and Recommended Practices, Annex 10 to Convention on Inter-national Civil Aviation, vol.1, Amendment 86, Nov. 2011.
10) 小西貞則，北川源四郎: 情報量規準（シリーズ・予測と発見の科学），朝倉書店, (2004).
11) 統計数理研究所: 統計数理, Vol.52, No.1（特集「極値理論」）, (2004).

森永 聡（正会員）
E-mail: morinaga@cw.jp.nec.com
1994 年東京大学大学院工学系研究科修了，日本電気株式会社入社．1999 年論文提出により学位（工学博士）取得．2000 年～2001 年金融監督庁出向，2004 年～2008 年金融庁兼務．データマイニングの研究・事業化に従事．

青木 健児（非会員）
E-mail: k-aoki@bq.jp.nec.com
2008 年北海道大学大学院情報科学研究科（博士）修了，日本電気株式会社入社．データマイニングの研究に従事．

鈴木 和史（非会員）
E-mail: k-suzuki@hq.jp.nec.com
2001 年同志社大学大学院工学研究科博士前期課程修了，日本電気株式会社入社．2009 年～2010 年スタンフォード大学客員研究員．GBAS 開発に従事．

藤巻 遼平（非会員）
E-mail: rfujimaki@nec-labs.com
2006 年東京大学大学院工学系研究科修了，日本電気株式会社入社．2011 年論文提出により学位（工学博士）取得．2011 年より NEC 北米研に出向中．データマイニングの研究に従事．

福島 荘之介（非会員）
E-mail: fks442@enri.go.jp
1994 年電気通信大学博士前期課程了．2007 年東京海洋大博士後期課程了．博士（工学）．独立行政法人電子航法研究所主幹研究員．衛星航法による精密進入着陸システムの研究に従事．測位航法学会理事．

投稿受付：2012 年 10 月 25 日
採録決定：2012 年 11 月 24 日
編集担当：桑名栄二（日本電信電話（株））