リサーチペーパー第 34 号 Research Paper No.34

公的統計の高度な二次的利用のための秘密計算技術の応用の研究

Study of Secure Computation Application for Advanced Utilization of Official Statistics

高橋 克巳

統計研修所客員研究官

NTT セキュアプラットフォーム研究所 主幹研究員

Katsumi Takahashi

SRTI Guest Researcher

Senior Research Engineer, Supervisor, NTT Secure Platform Laboratories

平成 27 年 11 月

November 2015

総務省統計研修所

Statistical Research and Training Institute (SRTI)

Ministry of Internal Affairs and Communications

論文受理日：平成 26 年 3 月 24 日

本ペーパーは、総務省統計研修所の客員研究官が、その責任において行った統計研究の成果を取りまとめたものであり、その内容については、総務省統計局又は統計研修所の見解を表したものではない。

本研究では、統計法（平成 19 年法律第 53 号）第 32 条の規定に基づき、科学技術研究調査に係る調査票情報を使用した。

公的統計の高度な二次的利用のための秘密計算技術の応用の研究

高橋 克巳
統計研修所客員研究官

概要

公的統計の活用を行うためには、調査票に含まれる個人や企業に関する機密情報を守るために、秘密の保護が重要である。本研究では、暗号技術をベースに情報の秘密を守りながら統計分析を行うことができる秘密計算技術に着目し、統計への応用を検討した。統計業務で暗号での保護が有意義である処理プロセスを確認し、秘密計算の統計へ応用するシステムアーキテクチャを確認し、さらに、秘密計算システムを統計研修所内の環境に構築し、科学技術研究調査に係る調査票情報を使用して評価を行った。以上の検討から、秘密計算技術が公的統計の安全を確保することと、高度な二次的利用に貢献できることを明らかにした。

キーワード：二次的利用、秘密の保護、暗号、セキュリティ、秘密計算

Study of Secure Computation Application for Advanced Utilization of Official Statistics

Katsumi Takahashi
SRTI Guest Researcher

Abstract

For the advanced utilization of official statistics, security is important. The application of the Secure Computation - which enables the statistics while the data kept secret in cryptographic way - is studied. After the survey of the process and the architecture of Secure Computation for statistics, Secure Computation System is developed in SRTI, tested with the Survey of Research and Development Data, and shown that can be contributed to the security and the use of official statistics.

Keywords: Utilization of statistics, Confidentiality, Cryptography, Security, Secure Computation

目次

図目次

表目次

1 はじめに

公的統計の利用、とくに二次的利用には制限があったが、60 年ぶりに改正された統計法（平成 19 年法律第 53 号、平成 21 年 4 月施行）では、一般からの委託に応じて既存の調査票情報から新たな集計表を作成・提供（オーダーメード集計）することや、匿名性を確保した調査票情報（匿名データ）の利用が可能になった。

さらに昨今の「ビッグデータ」ブームは、様々なデータから価値を発見して公益や各種の業務に役立てようというもので、これも公的統計利用への期待を広い意味で後押ししているだろう。

このように公的統計の高度な利用への期待は高く、利用制度の整備も進んできている。

しかしながら、公的統計の二次的利用に関する問題が存在しないわけではない。「統計データの二次的利用促進に関する研究会」では、平成 23 年度の報告書[*1]の中で、検討に当たっての考え方や守るべき原則として、二次的利用は統計目的に限るとする原則とともに、下記の通りニーズへの対応と秘密の保護が重要としている。この有用性と秘密の保護を両立する統計利用の環境の構築が必要である。

- 二次的利用は、ニーズに過不足なく対応することが望ましい。
- 二次的利用における有用性の向上と秘密の保護は、二律背反である。このバランスを確保するにあたり、国民・企業の信頼を損なわないようにするため、データを秘匿する措置、漏洩等を抑止するシステムやプロセスの安全性の保障などを確保する。

一方、個人情報やパーソナルデータ（個人に関するデータ）を分析して利用するニーズや機会が社会的に増えたことにより、従来よりセキュリティやデータベースの分野で研究されてきたプライバシー保護データマイニング (PPDM) [*2]とよばれる技術に注目が集まっている。PPDM は、パーソナルデータを正しく保護した上で、情報処理の手段を提

[*1] 45 ページ参考文献 [1] 参照

[*2] PPDM: Privacy Preserving Data Mining。類似のやや広めな概念に PET: Privacy Enhancing Technology がある。

供する技術である。PPDM には、統計学的な技法で匿名化してプライバシーを保護する方法や、暗号学的にデータを保護する方法がある。

本研究では、公的統計の高度な二次的利用のために PPDM 技術の応用の検討を行った。特に暗号技術の応用として最先端の技術である、暗号化したまま統計計算ができる秘密計算技術に注目し、秘密計算システムを統計研修所内の環境に構築し、科学技術研究調査に係る調査票情報を使用して評価を行った。

本報告書の構成は次の通りである。まず公的統計の二次的利用に関する課題について述べ、秘密の保護に係る暗号技術と PPDM 技術の現状を概観し、秘密計算技術の概要について述べ、統計の二次的利用に秘密計算技術を適用する方法を提案し、その評価結果を述べる。

なお、本報告書の対象は公的統計で扱うデータであるが、必要に応じて統計に限定しない個人情報の処理に言及する。後者に言及する場合は、一般的な名称としてパーソナルデータという語を用いる。パーソナルデータとは個人情報や（個人に関する）統計データを含んだ概念である。

2　公的統計の二次的利用に関する課題

2.1　統計法にみる課題

統計法[*3]は公的統計の体系的かつ効率的な整備及びその有用性の確保を図り、国民経済の健全な発展及び国民生活の向上に寄与することを目的としている。その中で利用促進及び統計調査の対象者の秘密の保護に関して次のように定めている[*4]。

- 行政機関との共同研究など高度な公益性を有する研究などに限り、調査票情報を提供することができる（調査票情報の提供）。
- 統計の研究や教育など公益に資するために使用される場合に限り、委託により統計を作成・提供することができる（オーダーメード集計）。
- 統計の研究や教育など公益に資するために使用される場合に限り、匿名データ[*5]を作成・提供することができる。

- 行政機関等は調査票情報等（調査票情報及び匿名データ）を適正に管理するために必要な措置を講じなければならない。
- 調査票情報等の提供を受けた者は同情報を適正に管理するために必要な措置を講じなければならない。
- 調査票情報を取り扱うもの及び提供を受けたものは守秘義務がある。
- 調査票情報等の提供を受けたものは、提供を受けた目的以外の目的で利用または提供してはならない。
- 上記業務の委託を受けた者にも準用する。

[*3] 45 ページ参考文献 [2] 参照

[*4] 統計法 32 条〜36 条（38 ページ参考 1）、統計法 39 条〜43 条（39 ページ参考 2）を参照

[*5] 本リサーチペーパにおいて匿名データとは、統計法第 2 条 12 項が定めるもの、すなわち一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む）ができないように加工したものをいう。一方、個人情報を「匿名化」した情報一般に関しては「匿名化」データと表記している。

このように、二次的利用の拡大と秘密の保持を同時に行うことが基本的な要件となっている。詳しくは、38 ページ参考 1 の統計法（平成 19 年法律第 53 号）からの引用を参照されたし。

2.2 「統計データの二次的利用促進に関する研究会」にみる課題

総務省では、統計データの二次的利用等に関する検討を行うに当たり、利用者側からの意見等を反映させるとともに、技術的助言を得るために「統計データの二次的利用促進に関する研究会」を開催している。

この研究会の平成 23 年度の報告書では、諸外国を調査し二次的利用提供データの種類と提供方法の類型化、二次的利用に関する民間の意見、二次的利用推進に向けた取り組みの方向性をまとめている。

諸外国の提供データは大きく 3 種類に分類されるとしている。

調査票情報レベルのデータ

調査客体を直接識別できる項目（名前、住所等）を削除した程度の秘匿処理を行った機密性の高いデータ。

匿名データ（加工処理の度合いが低いデータ）

個人・法人等を間接的に特定できる項目の削除、一定以上あるいは一定以下の回答をまとめて表示するトップ（ボトム）コーディング、特定の項目を一つのグループでまとめるグルーピング等の秘匿処理を行ったデータ。個人情報等の漏えいリスクは低い。（※日本の匿名データは、ここに分類されると考えられる。）

パブリックユースファイル（強度の加工処理を行ったデータ）

上記に加え、攪乱値を挿入するパータベーション、一部のデータを入れ替るスワッピングなど、程度の高い秘匿処理を行ったデータ。個人情報等の漏えいリスクはほぼない。（※利用目的に特段の制限なし）

データの提供方法については、直接利用型、プログラム送付集計型及びオーダーメード型の 3 種類の類型に分類でき、細分化すると 6 種類の類型に分類できるとしている。

表 1　提供方法の類型

1.　直接利用型	1.　オンサイト型	調査実施者の指定する施設内において、調査実施者が提供するデータを利用することができる。
	2.　直接提供型	利用者の研究室など、特定の施設以外の場所で、一定の条件のもとに調査実施者が提供するデータを利用することができる。
2.　プログラム送付集計型	1.　参照可能型	ネットワークを経由して調査実施者が提供するデータを閲覧した上で、作成したプログラムを送信し、集計された結果を受け取ることができる。
	2.　参照不可能型	ネットワークを経由して調査実施者が提供する参考情報（ダミーデータ等）を閲覧した上で、作成したプログラムを送信し、集計された結果を受け取ることができる。
3.　オーダーメード型	1.　従来型（後日提供型）	調査実施者に対して集計の委託を行い、調査実施者によって集計された集計結果表を受け取ることができる。
	2.　リアルタイム提供型	インターネットを介してシステムにアクセスし、集計項目を指定することによって、自動集計された集計結果表をリアルタイムに受け取ることができる。

この中で、我が国で提供されているのは、表 1 の 1-2 と 3-1 にあたる以下である。

- 直接利用型（直接提供型）による調査票情報の提供、匿名データの提供。
- オーダーメード型（後日提供型）による匿名データの提供。

これに対して、同報告書では民間等の意見として以下が上げられている。

- オーダーメード集計に関して、あらゆるクロス集計を行いたいので、ローデータの提供を受けたい。
- オーダーメード集計で結果を得るまでの期間を短縮してほしい。

- 高度の匿名化がなされたパブリックユースファイルは、ビジネスでは使わない。
- 匿名データを利用するセキュリティ上の利用条件が厳しい。

以上から、提供情報の二次的利用の拡大のニーズがありながら、一方で解決すべき技術上の問題、秘匿上の問題、制度上の問題があるとしている。

2.3 秘密の保護の観点からみる課題

秘密の保護は、いわゆる情報セキュリティ及びプライバシー保護の両方の観点から整理することができる。情報セキュリティとは情報の機密性・完全性・可用性を守ることで、統計では調査票データそのものを保護する行為である。一方、プライバシー保護とは調査票データに含まれる個人のプライバシーを確保することで、パーソナルデータの取り扱いの取り決めを明確かつ誠実に行うことと考えることができる。統計では、目的、範囲、事項、方法などから構成される調査行為と、集計事項、公表方法などの結果を求める行為が取り扱いに含まれる。

後者のパーソナルデータの取り扱いに関して、留意点をまとめたものに、OECD の 8 原則[*6]がある。このガイドラインは個人情報の国際流通について述べたもので、日本の個人情報保護法を始め、各国の個人情報保護に関する法律の基本的な構成要素になっている。(掲載順及び注釈は筆者による。)

OECD のプライバシー保護と個人データの国際流通についてのガイドライン

1. 目的明確化の原則
   収集目的を明確にし、データ利用は収集目的に合致するべき。
2. 利用制限の原則
   データ主体の同意がある場合、法律の規定による場合以外は目的以外に利用してはならない。
3. 収集制限の原則
   適法・公正な手段により、かつ情報主体に通知又は同意を得て収集されるべき。
4. データ内容の原則
   利用目的に沿ったもので、かつ、正確完全、最新であるべき。
5. 安全保護の原則

[*6] 45 ページ参考文献 [3] 参照

合理的安全保護措置により、紛失・破壊・使用・修正・開示等から保護するべき。

6. 公開の原則
   データ収集の実施方針等を公開し、データの存在、利用目的、管理者等を明示するべき。
7. 個人参加の原則
   自己に関するデータの所在及び内容を確認させ、又は異義申立てを保証するべき。
8. 責任の原則
   管理者は諸原則実施の責任を有する。

ここで述べられている要件は、利用目的を明確にして同意をとってデータの収集を行うべきということである。一般的に公的統計の基本的業務はこの基本的な要件を満たしているだろう。高度な二次的利用を行うために留意すべきことはなんであろう。

OECD のガイドラインと並んで、プライバシー保護のために参照される考え方にプライバシー・バイ・デザインがある。カブキアン博士は同名の著書[*7]の中で、収集制限について次のように定義し、取り扱うデータを最小限にすることを提唱している。

> 個人情報の収集は公正で合法的な方法で行い、明示された目的を達成するために必要なものに限定されていなければならない。
> データは最小限に ― 個人情報の収集は、厳密に最小限に留めなければならない。プログラム、情報技術、システムの設計では相互作用及び取引から個人を特定できないことを初期条件とする必要がある。可能な場合は、個人情報の特定可能性、観測可能性、リンク可能性を最小限にする。

[*7] 45 ページ参考文献 [4] 参照

2.4 課題のまとめ

以上から、公的統計データの二次的利用に関する課題を以下の通り認識して本研究の前提とした。

- 公的統計の二次的利用の拡大と秘密の保護の両立が基本的な期待である。
- 二次的利用では詳細な分析、特に自由度の高いオーダーメード集計への期待が大きい。
- 二次的利用の推進のためは、個人情報の安全な管理が重要。

3 暗号技術及びプライバシー保護データマイニング技術の現状

3.1 パーソナルデータ処理のプロセス

プライバシー保護データマイニング技術を理解するために、先に述べた OECD の 8 原則に加え、日本の個人情報保護法[*8]、カブキアン博士の “Global Privacy Standard”（「プライバシー・バイ・デザイン」所収）などの代表的な法やガイドラインからパーソナルデータへの要件をまとめてみる。

パーソナルデータの取り扱いでは、誰から誰に、どのようなパーソナルデータが流通するのかを明確に意識する必要がある。このモデルでは、秘密にすべき情報を提供する「個人」と、そのデータを取得する「取得保有者」と、そのデータの提供を受け利用する「活用者」の 3 つのエンティティを仮定する。統計の場合は、調査票に回答する個人、統計局、研究者がそれぞれのエンティティの例である。ネット通販履歴処理の応用を考えた場合、ネットユーザ、ネットショップ、マーケティング分析会社がそれぞれのエンティティに相当する。複数のエンティティが同一、例えば取得保有者と活用者が同一のエンティティでもかまわない。図 1 にパーソナルデータ処理のプロセスモデルを示す。

自分の情報が、知らないうちに、本来の目的以外の理由で利用されることは、快いことではなかろう。これが最も重要な原則「目的の特定」と「同意の取得」である。この原則は、パーソナルデータの取り扱いの最初から最後まで配慮されるべきである。

それに伴い、取得されるパーソナルデータは、目的の遂行に必要最小限のデータであるべきであり、「セキュリティ」を確保して安全に保有されなければいけない。

さらに、パーソナルデータの活用を考えてみる。取得保有の段階であれば、そのデータに接触する人や機会を制限することが比較的やりやすいだろうが、活用の段階では、活用されるデータが多くの人の目につく可能性もある。その活用状況から、活用されるデータに含まれる個人のことがいろいろ詮索される状況は排除するべきである。すなわち、パーソナルデータの活用においては、必要がないのであれば、できるだけ個人にひもづいたデータは使うべきではなく、「匿名性」が保証されたデータを使うべきである。

[*8] 45 ページ参考文献 [5] 参照

3.2 パーソナルデータの分類

パーソナルデータの取り扱いで匿名性が重要な概念であることを述べた。これを理解するために、パーソナルデータを形式的に分類する。

1. 実名データ
   実名データとは、個人を直接識別することができる氏名、生年月日などの記述を含むデータ。個人情報保護法の「個人情報」に該当し、利用には多くの制約がかかる。
2. 「匿名化」データ[*9]
   「匿名化」データとは、実名データを個人が特定できないように加工したデータ。
   - 「単純な匿名化」データ
     実名データから氏名、生年月日などの個人を識別することができる記述がないように加工したもの。
   - 「高度な匿名化」データ
     単純な識別情報の削除だけでは、個人が特定される可能性があるため、さらに属性値を曖昧にしたり希少な属性を削除するなどの処理を施したもの。
3. 統計化したデータ
   統計化したデータは、同じ性質を持った個人を数え上げたデータ（表）。統計化したデータは一般に高い抽象度で数え上げを行うため、「匿名化」データよりも、より個人識別性に関して安全であると考えられる。

[*9] 本リサーチペーパにおいて「匿名化」データと表記する場合は、個人情報を「匿名化」した情報一般に関するものとし、統計法における匿名データ（3 ページ脚注参照）を限定する場合と使い分ける。

図１　パーソナルデータ処理のプロセス

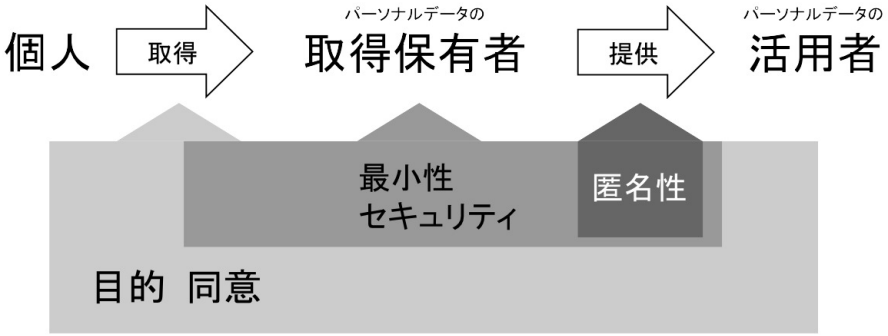


図２　パーソナルデータ処理の分類

1. 実名データ

氏名	生年月日	勤務地	趣味
鈴木二郎	1973.10.23	東京都千代田区緑町1	野球
三浦数良	1967.02.27	埼玉県浦和市一番町2	サッカー

2. 単純な匿名化データ

氏名	生年月日	勤務地	趣味
鈴木二郎	削除	東京都千代田区緑町1	野球
三浦数良	1967.02.27	埼玉県浦和市一番町2	サッカー

2'. 高度な匿名化データ

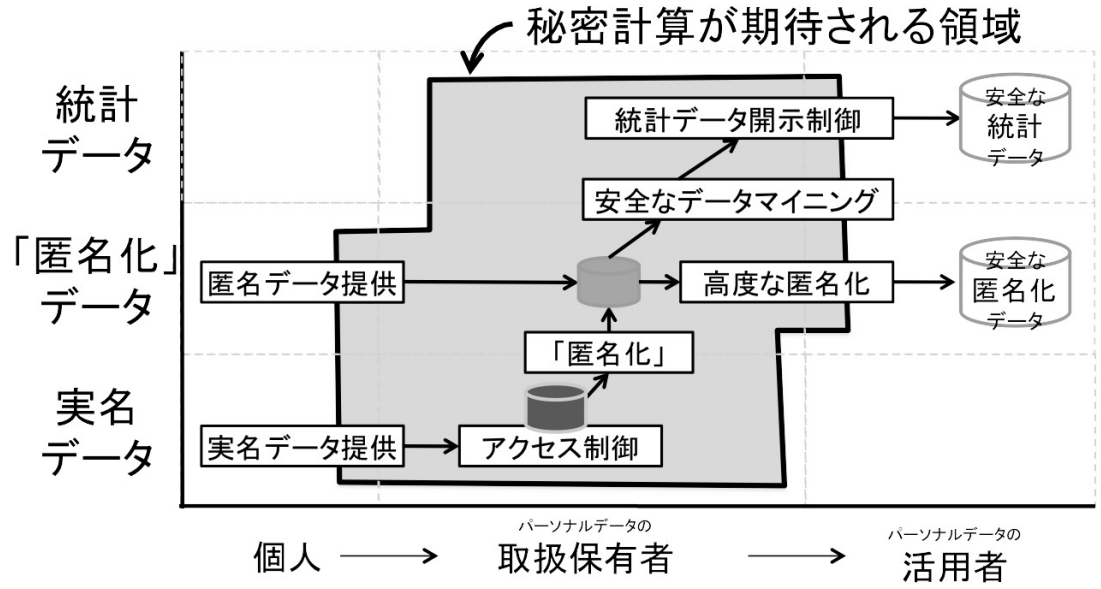
勤務地	趣味
東京都	球技
埼玉県	球技

3. 統計データ

	東京	埼玉
野球	33	8
サッカー	27	32

3.3 暗号技術及びプライバシー保護技術の分類

図 3　プライバシー保護技術の分類



暗号技術及びプライバシー保護データマイニング（PPDM）の概要及び秘密計算の PPDM における位置づけについて説明する。

図 3 に暗号技術及びプライバシー保護データマイニング（PPDM）の要素技術の分類を示す。PPDM の目的の直感的理解は、プライバシーを保護するデータを、データを「取り扱い保有する」組織や人や「活用をする」組織や人から、「高度な匿名化データ」や「統計データ」に変形することによって守ることである。なお、PPDM の全体像に関しては、サーベイ文献を参照のこと[*10]。

3.4 匿名化

データを暗号化しないで保護するためには、データを変形または削除させることにより、見られると困る部分を隠すことになる。この操作を「匿名化」と呼ぶ。

匿名化技術の PPDM のシナリオとしては、実名データの「高度な匿名化」による安全な匿名化データの作成となる。個人から実名データが提供されたとき、取扱保有者の役割

[*10] 45 ページ参考文献 [6][7] 参照

はアクセス制御などを行ってそのデータを適正に管理し活用者に安全なデータを提供することであり、活用者はそのデータを使って、自分で最終目的のデータ分析処理を行う。

匿名化技術には複数の方法が知られている。これらの方法を適切に組み合わせることによって、データを活用するために第三者に提供を行う際にも秘密の保護を行うことが可能になる。なお暗号と異なり、「匿名化」は単純に施したからといって、必ずしもデータが保護できる訳ではない。「匿名化」の基本的な考え方は、まず氏名等の個人が直接識別できる属性は削除し、年齢や勤務先など個人が間接的に識別できる情報はあいまいにし、保護の必要がない属性だけをオリジナルのままで使うというものになる。

図 3 の実名データとは個人が識別できるデータである。例えば氏名や電話番号などの個人が直接識別できる属性を含んだデータであるが、この実名データを個人のプライバシーに関する問題がない状態に変換して「安全な匿名化データ」にすることが、匿名化の本質的な目的である。

匿名化には様々なレベルがあることに注意されたい。図 3 には「匿名化」及び「高度な匿名化」と記した。前者は実名データから単純に個人識別性のある属性を取り除く、あるいは元の属性値と結びつきがない値に変換する作業である。例えば、対象のデータが氏名及び身長と体重の 2 種の属性からなる場合、氏名を削除してしまえば、そのデータはそれ単独で個人を特定することができなくなる。しかし、属性数が少数ではなく多種になる場合は、個人が特定できる可能性を完全には排除することができない。このためには「高度な匿名化」を行って、個人が特定される可能性を少なくすることを含め、影響を最小化することが求められる。「高度な匿名化」には様々なテクニックが用いられるが、例えば属性値をあいまいにしておくことで（例、身長、体重の値を四捨五入する）、その属性を組み合わせても対象となる人が複数人いるようにデータを変形する。複数人を「k 人以上」とパラメータ表記する場合、これに対応する匿名化は「k-匿名化」と呼ばれる[*11]。また、属性に現れる頻度の少ないもの（例、希少疾病）やそれが漏れることで影響が大きいセンシティブなもの（例、思想信条）が含まれる場合は、それらの属性を削除するといった対処も用いられる。

匿名化の代表的な手法を下記に示す。

[*11] 45 ページ参考文献 [8] 参照

識別情報の削除

氏名などの直接識別できる情報を削除する。

保護属性の削除

特に大きい、もしくは小さい特殊な属性を持つ個人のデータをレコードごと削除する方法。例えば、100 歳以上の人のレコードは削除する。

保護属性のトップ（ボトム）・コーディング

特に大きい、もしくは小さい特殊な属性をまとめる方法である。例えば、101 歳や 110 歳の人は「100 歳以上」とする。

保護属性のグルーピング

年齢を 2 歳刻みあるいは四捨五入する、住所を市区町村で扱うあるいは都道府県で扱うなどの属性値の一般化を行ってグループ化する。

保護属性に対する誤差の混入やスワッピング

数値属性に誤差を混入したり（年齢 20 歳を 21 歳にしてしまう）、カテゴリー属性を入れ替える（A さんの住所と B さんの住所を入れ替えてしまう）。

匿名化したデータの並び替え

例えば、アルファベット順に並んだデータの並びからの元データへの推測を防ぐためにデータの並び順を変える。

匿名化したデータのリサンプリング

データを全て提供するのではなく、確率的に選択（サンプリング）し、提供データと元データの関係を一意にならないようにする。

保護が必要なレコードの削除

上記の方法を行っても、レコードが一意に特定でき、そこから個人の識別が可能になる場合は、そのレコードを他に提供する時点で削除する。

3.5 安全なデータマイニング

「安全なデータマイニング」には暗号技術を使うものや、機械学習技術を使うものがある。暗号技術を使うものは秘密計算が代表的である。秘密計算には数多くのバリエーショ

ンが用途に応じて存在する[*12]。機械学習技術を使うものには撹乱再構築法と呼ばれるものがあり、データに非可逆な確率的な操作を施し元のデータの推定を困難にし（撹乱）、さらに撹乱したデータから元のデータの分布を推定して、元データと同様の統計的性質を持った擬似データを作成する（再構築）。この擬似データを使うことで、マイニングが安全になる。撹乱の方法にはノイズの付加や、値同士のスワッピングなどがある。再構築の方法にはベイズ推定を用いる方法がある[*13]。秘密計算に関しては、後で詳述する。

3.6 統計データ開示制御

もう一つの PPDM の代表的なシナリオは、データマイニング結果からプライバシーを漏らさないことである。実名データや「匿名化」データが提供されたとき、取扱保有者の役割はデータマイニングまで行って、そのマイニング結果を安全な統計データであることの確認を行った上で活用者に提供することである。活用者は「PPDM システム」と対話を行いながら、マイニング業務を行う。このシナリオにおいてデータを取扱保有者から保護する機能が「安全なデータマイニング」、データマイニングの結果の安全さを確認する機能が「統計データ開示制御」である。

「統計データ開示制御」は、いわゆる統計分野で永年研究されてきた蓄積がある。基本的な考え方は、開示する統計値のうち、特定の個人の影響が強いもの（例えば、統計結果が一名の個人のデータしか反映しないもの）を見つけ出し、それを削除するものである[*14]。この手法のほかに、統計値に対してノイズを付与してプライバシーを保護する技術の研究も知られている[*15]。

3.7 暗号

暗号技術はセキュリティとデータを最小限にすることに貢献する。暗号化されたデータは、権限（復号鍵）を持たないものからは閲覧が不可能であるため、情報漏洩対策等のセキュリティに貢献する。さらに閲覧を制限することによって、特定可能性、観測可能性、リンク可能性[*16]などのデータを最小限にする原則に貢献する。現在暗号ではデータや通信の秘匿と改竄防止を軸に、様々な機能が実現されている。

[*12] 45 ページ参考文献 [9] 参照
[*13] 45 ページ参考文献 [10] 参照
[*14] 45 ページ参考文献 [11] 参照
[*15] 45 ページ参考文献 [12] 参照
[*16] 7 ページ参照

関数型暗号

関数型暗号（Functional Encryption）はロジック（関数）を暗号データもしくは鍵に定義することができる公開鍵暗号の一種である。例えば、暗号データに復号の条件をロジックで定義（例、○○部署かつ管理職のみが閲覧できる）した場合、その属性を鍵として正当に保有するもののみが、そのデータの復号が可能である。この技術を使うことにより、データを暗号化した上で自由に流通させ、特定のシステムを離れても、データの閲覧の制限ができることが期待されている。

準同型暗号

暗号の準同型性（Homomorphism）とは暗号が数学的構造を保つことである。例えば、公開鍵暗号や秘密分散には乗法準同型性や加法準同型性を持つものがあり、暗号化されたままの乗算や加算が可能である。これらを性質を使って、データを委託しても委託先でデータの平文を閲覧されることなしの計算処理ができることが期待されている。この性質に基づいて暗号プロトコルとして実装されたものが秘密計算と呼ばれる。

秘密分散

秘密分散（Secret Sharing）は秘密にする情報を、複数に分けることによって保護するアルゴリズムである。分散した情報を一定数集めることによって、情報を復元することができ、逆に一定数集めない限り、秘密が漏れることはない。秘密分散には準同型性を持つものがある。

秘密計算

秘密計算（Secure Computation）は準同型暗号や秘密分散などの暗号技術、暗号プロトコルに基づいて、データが秘匿されたまま各種の計算を行うことができるシステムである。入力データは秘匿されて秘密計算システム配置される。秘密計算システムを構成するコンピュータは、データの中身を閲覧することなしに、あらかじめ定められた手順に従ってデータ処理を行い、最終的に求める計算結果を、秘匿された形式で出力する。秘匿された結果を復元することで最終結果を得る。様々な形態を持ち、秘匿関数計算（Secure Function Evaluation）、セキュアマルチパーティー計算（Secure Multiparty Computation）などと呼ばれるものもある。

3.8 秘密計算

最後に紹介する PPDM の技術が秘密計算である。秘密計算の役割はデータを暗号化して、データの処理プロセスにおける漏洩や不正な閲覧といった様々なリスクを回避しながら、最終的に実名データを安全な統計データに加工することである。狭義の秘密計算は、図 3 における「安全なデータマイニング」の領域であり、マイニング時のリスクを回避する。さらに暗号化する領域を個人の側に拡張することによって、実名データが個人のところで作成する時点からの保護を開始することができる。また、例えば以下のようなシナリオも考えられ、対応した研究開発が期待される。

- 「秘密計算」と「統計データ開示制御」と組み合わせることにより、データ分析プロセスからも分析結果からもプライバシー問題を回避する。
- 「秘密計算」で「高度な匿名化」を行えば、「匿名化」データの作成プロセスを安全にする。

秘密計算技術の研究開発は世界各国で行われており、著者らの NTT 秘密計算プロジェクト[*17]の他に、エストニアの Sharemind[*18]や、デンマークの VIFF [*19] が知られている。NTT 方式は医療統計への応用を報告しているが[*20]、Sharemind は企業財務データ分析、VIFF はオークションなどをターゲットに研究を進めている。

各方式を比較すると、基本処理性（乗算回数）は NTT が 100 万回/秒、Sharemind が 100 万回/秒、VIFF が 300 回/秒などの値が報告されている。各研究機関とも実証実験を行いながら、商用実用化を急いでいる。

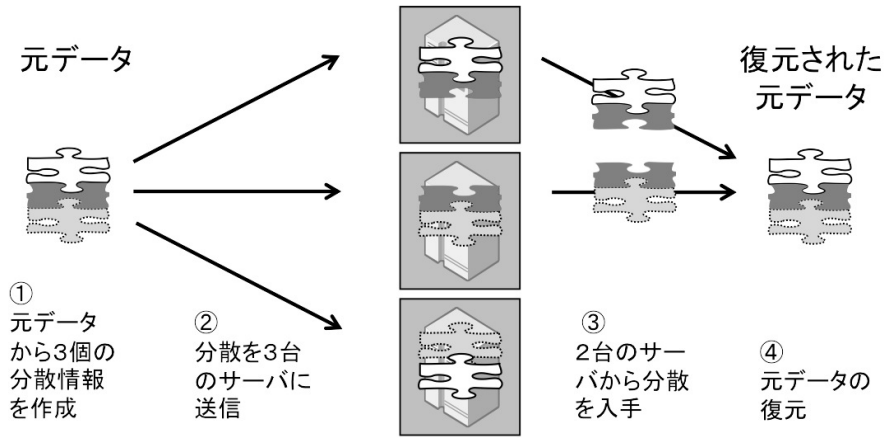
[*17] 46 ページ参考文献 [13][14] 参照
[*18] 46 ページ参考文献 [15] 参照
[*19] 46 ページ参考文献 [16] 参照
[*20] 46 ページ参考文献 [17] 参照

4　秘密計算技術の概要

前章で述べた、暗号の準同型性を用いてデータを暗号化したまま分析処理を行う技術に秘密計算技術がある。本研究では、NTT セキュアプラットフォーム研究所で開発が行われている秘密分散に基づく秘密計算のプロトタイプシステムの統計業務への適用を検討する。なお、以降秘密計算と表記した場合は、NTT 秘密計算をさすものとする。

4.1　秘密分散

図 4　秘密分散



秘密計算技術のデータの機密性は、秘密分散を基本としている。

秘密分散は秘密にする情報を、複数に分けることによって保護するアルゴリズムである。分散した情報を一定数集めることによって、情報を復元することができ、逆に一定数集めない限り、秘密が漏れることはない。

図 4 は、情報を 3 つに分散させ、2 つを集めることによって復元できる秘密分散の例である。

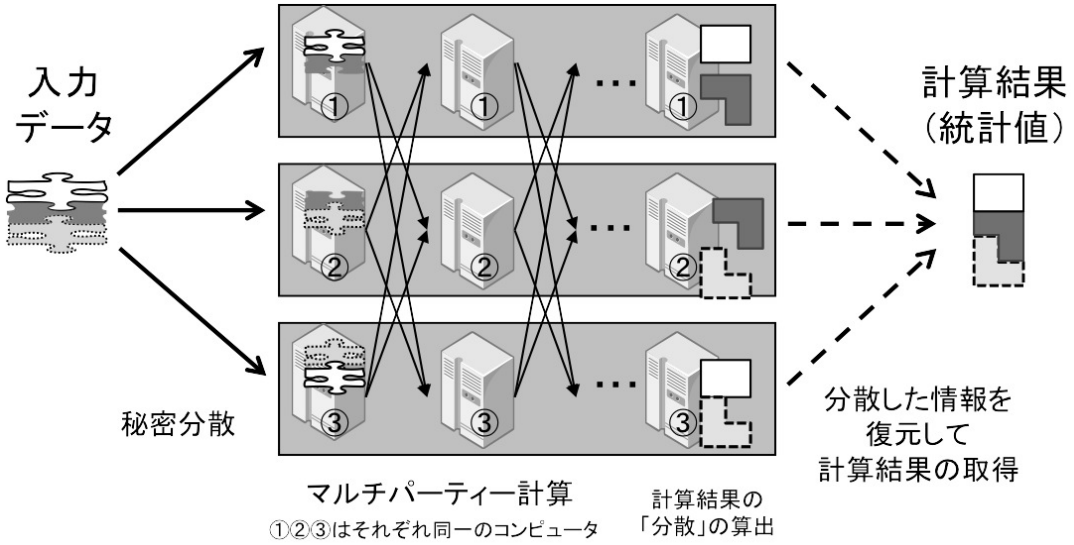
Shamir の秘密分散[*21]は多項式補間問題に基づく方法で、全体の分散数を n、復元に必要な分散数を k とすると、その多項式上の点 n 個から、k 個を集めて k-1 次多項式を決めることである。この n 個の分散した情報は情報理論的安全性を持つ。情報を n 台のコンピュータに分散させることによって、秘密分散システムが構築できる。

[*21] 46 ページ参考文献 [18] 参照

- データの保存：データを複数に分け n 台のコンピュータに分散させ保管。
- データの復元：k 台のコンピュータから分散した情報を集めデータを復元。
- 機密性：k-1 台までのコンピュータからデータを盗んでも何の情報も得られない。
- 可用性：n-k 台まで故障しても、残りのコンピュータからデータを復元できる。

4.2 秘密計算

図 5　秘密計算の概要



秘密計算は、秘密分散の考えを情報の安全性のベースにし、データが秘匿されたまま各種の計算を行うことができるシステムである。図 5 は、データを秘匿して計算処理を委託する委託型秘密計算システムの概念図である。入力データは秘密分散されて複数のコンピュータに配置される。秘密計算を構成するコンピュータは、分散したデータの中身を閲覧することなしに、あらかじめ定められた手順に従ってデータ処理を行い、計算結果を結果の値の分散した情報として計算出力する。最終的な結果の出力は秘密分散と同様に、必要な数の分散した情報を集めることによって行われる。

- データの保存：元データを複数に分け n 台のコンピュータに分散させ保管。
- マルチパーティー計算：行うデータ処理に応じて、コンピュータ間で必要なデータ処理とデータの送受信（通信）を必要回数行う。この際、各コンピュータにおいて、他から受信したデータと各コンピュータが保有する分散した情報を合わせても各コ

ンピュータは何も元データに関する情報を得られない。

- 機密性：計算過程において各コンピュータは何の情報も得られず、k-1 台までのコンピュータからのデータが集められても何の情報も得られない。
- 結果の出力：計算結果が各コンピュータに分散した情報として出力される。

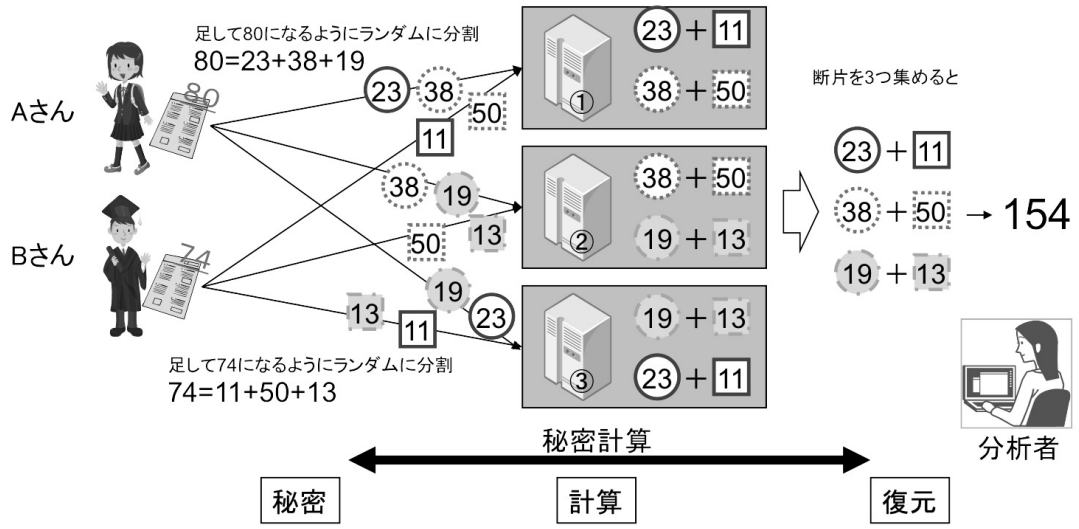
秘密計算システムは、上記のマルチパーティー計算環境上に論理演算や算術演算を構築することにより、各種、統計を含む計算が可能になる。

4.2.1 秘密計算の加算

秘密計算システムは、独特の方法で算術演算を実現する。まず始めに秘密計算による加算の実現方法を説明する。

図 6 は秘密計算で 2 つのテストの点（整数）の足し算を行う概念を説明する例である。ここではテストの点は適当な数字（乱数）を使って、3 つの数の和に分散され、秘密分散される。各コンピュータでは、各々が分散データを使ってローカルな加算器として振る舞い、全体の足し算の秘密分散を得る。なお、本秘密計算が使う秘密分散はこのような整数の和で表されるものではないが、基本的な概念は同じである。ここで分散値の和が最終的に全体の和になる性質は、秘密分散の加法準同型性という性質が用いられている。

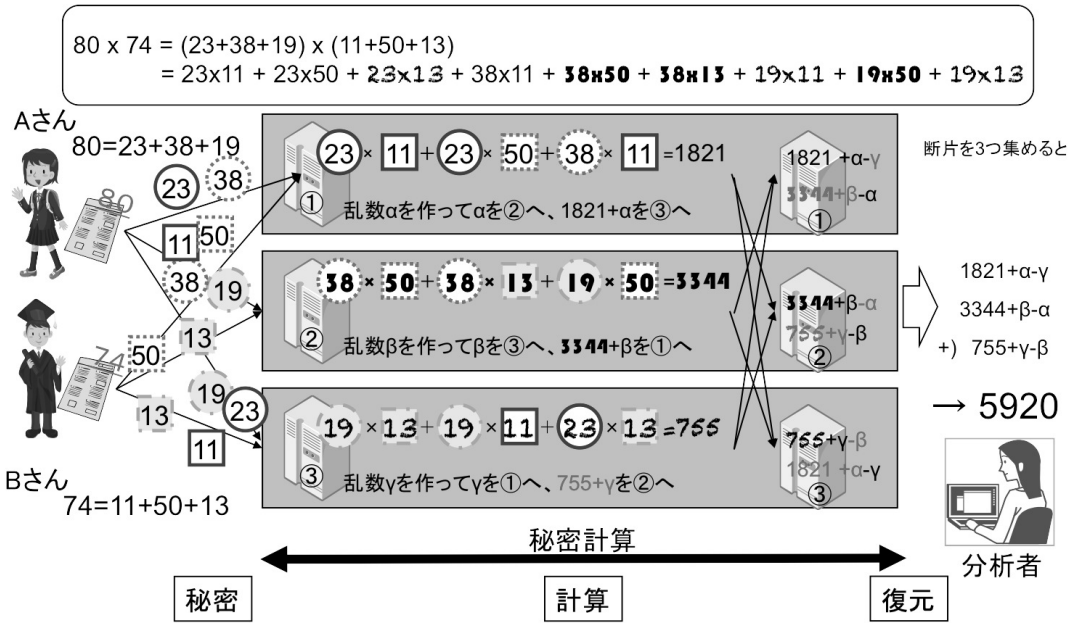
図 6　簡単な足し算の例



4.2.2 秘密計算の乗算

続いて秘密計算システムにおける乗算の実現方法を説明する。乗算は加算のテクニックを拡張して行う。2 つの整数の積は、図 7 図頭に示した、3 つの整数同士の積の和として考えることができる。この積の和の分散値をマルチパーティー計算で求めることが目的となる。図に示した通り、求める部分的な積の和は各コンピュータ内で計算することができる。この積の和を積の秘密分散の形で各コンピュータに配置させるためには、乱数を使いデータ秘匿を行った通信を行う。以上で乗算が実現できる。

図 7　簡単なかけ算の例



4.2.3 秘密計算の論理演算

秘密計算システムにおける論理演算は、この加減算と乗算を用いて実現する。

論理積、論理和、排他的論理和は $a, b \in 0, 1$ を秘密分散させ、a, b に対する秘密計算の足し算と掛け算を行うことで実現できる。

- 論理積：$a \wedge b = ab$; なぜなら a=1 かつ b=1 の時のみ 1 となり、他は 0 となる
- 論理和：$a \vee b = a + b - ab$; なぜなら a=0 かつ b=0 の時のみ 0 となり、他は 1 となる
- 排他的論理和：$a \oplus b = a + b - 2ab$; なぜなら a=1 または b=1 の時のみ 1 とな

り、他は 0 となる

また、否定は $a \in 0, 1$ を秘密分散させ、秘密計算の足し算を行うことで実現できる。

- 否定：$\neg a = 1 - a$; なぜなら a=1 のとき $\neg a = 0$、a=0 のとき $\neg a = 1$

これらの論理演算を組み合わせることで、原理的に任意の計算処理が可能になる。

4.2.4 秘密計算システムの提供する演算

秘密計算システムでは、以下の演算を提供している。

- 平均値、分散値、最大値、最小値、中央値
- 条件によるフィルタリング
- カテゴリ別統計値 (集計、平均値、分散値、最小値、最大値、中央値)

これらの演算は上述の通り、算術演算や論理演算の組み合わせにより作ることも可能である。しかし、秘密計算システムでは算術演算や論理演算に加えて、ソートや比較などの個別に設計したより高度な演算を組み合わせてこれらの演算を作っている。

その理由は、算術演算や論理演算に基づいた一般的な構成方法で作ると多くの演算で処理性能が低くなってしまうからである。秘密計算では、動作の違いによって本来明かしてはならないデータについての情報が漏れることを防ぐため、入力や計算途中のデータに応じて動作を変えることは許されない。このため、一般的な構成方法では、情報を漏らさないために、条件分岐があった場合は両方の場合を計算して本来の分岐先の結果だけを残すという処理が行われる。その結果、条件分岐が多く含まれる演算を一般的な構成方法を使って秘密計算で実現すると計算時間が非常に大きくなってしまう。そこで、秘密計算ではいくつかの重要な演算を個別に効率よく実現し、これらを部品として用いて演算を実装している。

4.2.5 秘密計算のソート

秘密計算システムで使われている個別に設計された演算の一つにソートがある。ソートは入力として受け取った値の列を昇順に並び替える演算であり、通常の計算機上でも多くの処理の部品として使われる重要な演算である。秘密計算システムにおいても、最大値や最小値、中央値、カテゴリ別統計値の計算において、ソートは中心的な役割を果たす。秘密計算におけるソートは、秘密分散された値の列を入力とし、別の秘密分散された値の列を出力する処理であり、出力を復元した列が入力を復元して昇順に並べ変えた列と一致す

るように計算される。

秘密計算システムでは、公開しても入力に関する情報が一切漏れない中間データを作りだし、この中間データに依存して動作を変える工夫により、一切動作を変えないために非常に計算時間が大きくなってしまう一般的な構成方法に比べて効率よくソートを実現している。具体的には、以下の手順で処理を行う[*22]。

1. 各要素のソート後の順位を計算
2. 全体をランダム置換
3. 順位だけを復元
4. 復元した順位に従って移動

ソート後の順位の計算は、算術演算と論理演算の組み合わせにより効率よく行われる。ランダム置換[*23]は、秘密計算用に個別に設計された演算であり、秘密分散された値の列を入力として、出力の秘密分散された値を復元した列が入力の秘密分散された値を復元してランダムな順序に並び替えた列となっているように、別の暗号文の列を計算して出力する処理である。ランダム置換を用いて誰にもわからない順序に並び替えることにより、計算された順位を公開しても入力に関する情報は一切漏れない。最後に順位を復元し、これを利用して昇順への並べ替えを実現する。このように一般的な構成方法では必要となる条件分岐を回避したことで、ソートが効率よく実現される。

4.2.6 秘密計算の計算時間

次に、秘密計算の計算にかかる時間について説明する。まず秘密計算加算にかかる時間は、1 回の加算のために 2 回の加算器を動かす（図 6）が、この加算器が行う処理は、素数 p を法とする素体上で行われる特殊な加算が行われる（剰余加算）。同じ処理が 3 台のコンピュータで行われるが、並列に行われているため、システム全体の処理時間には 1 台で剰余加算を 2 回行う時間を考えれば良い。秘密計算加算に必要な時間は、剰余加算にかかる時間を T_{ModAdd} とすると以下である。

$$T_{SecAdd} = 2 \times T_{ModAdd} \tag{1}$$

秘密計算乗算の計算時間は、積の和を求めるためにそれぞれのコンピュータは積算器を 3 回、加算器を 2 回、乱数生成を 2 回とデータ通信、さらにデータ交換後に加算器を各 4

[*22] 46 ページ参考文献 [19] 参照
[*23] 46 ページ参考文献 [20] 参照

回使う必要がある（図 7)。

秘密計算乗算に必要な時間は、剰余加算にかかる時間を T_{ModAdd} 、剰余乗算にかかる時間を T_{ModMul} 、乱数生成にかかる時間を $T_{GenRand}$、通信にかかる時間を T_{com} とすると以下である。

$$T_{SecMul} = 3 \times T_{ModMul} + 2 \times T_{ModAdd} + T_{GenRand} + T_{com} + 4 \times T_{ModAdd} \quad (2)$$

参考に NTT における実装の報告値を紹介する。秘密計算加算 T_{SecAdd} 1 回に必要な処理時間の理論値は約 20 ナノ秒、秘密計算乗算 T_{SecMul} 1 回に必要な処理時間は約 1000 ナノ秒である。乗算の値を比較すると、秘密計算を行わない場合と比べて 10000 倍以上のオーバヘッドがあるといえる。

表 2　秘密計算処理の基本要素処理時間 (単位:ナノ秒)

剰余加算 T_{ModAdd}	剰余乗算 T_{ModMul}	乱数生成 $T_{GenRand}$	通信時間 T_{com}	秘密計算加算 T_{SecAdd}	秘密計算乗算 T_{SecMul}
10	250	70	100	20	980

一般の計算処理を論理回路 で行う場合は、上記の加算、乗算を最少単位として構成した論理回路の規模に応じた計算コストが必要である。

実証における基本性能は、平均値 0.79 秒、分散値 0.81 秒、中央値 4.95 秒、集計（1 群）4.86 秒（以上、1000 件の臨床データでの実証値)。ソート 20 秒（10 万件のデータ）である (Intel Core i7 2.7GHz × 2 コア、LAN 環境で測定)。[*24]

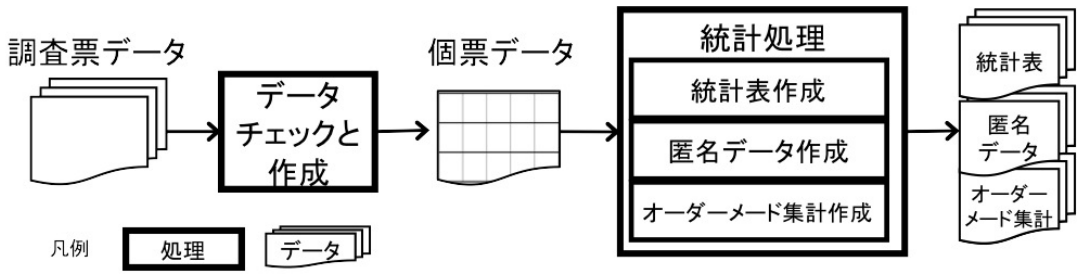
[*24] 上記の処理時間に関する数値は前出 46 ページ参考文献 [17] に基づく

5　統計の二次的利用のための秘密計算の応用

5.1　統計業務のモデル化

統計の一般的な業務プロセスを分析したところ、図８の通りであることが分かった。調査票データは記述内容のチェックが行われた後、テキスト表形式の個票データにまとめられる。次に個票データを用いて、各種の統計表が作成され、さらには二次的利用（オーダーメード集計や「匿名化」データ作成等）にも用いられる。

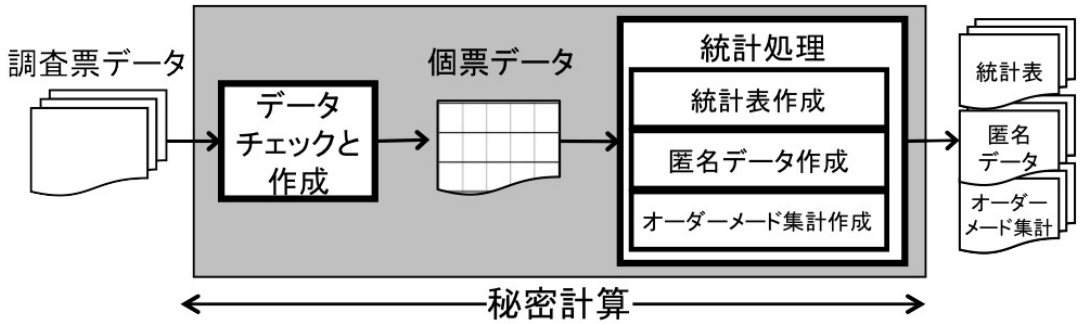
図８　一般的な統計の業務プロセス



この統計の業務プロセスに対して、秘密計算技術を使って秘密の保護を行うことを考える。

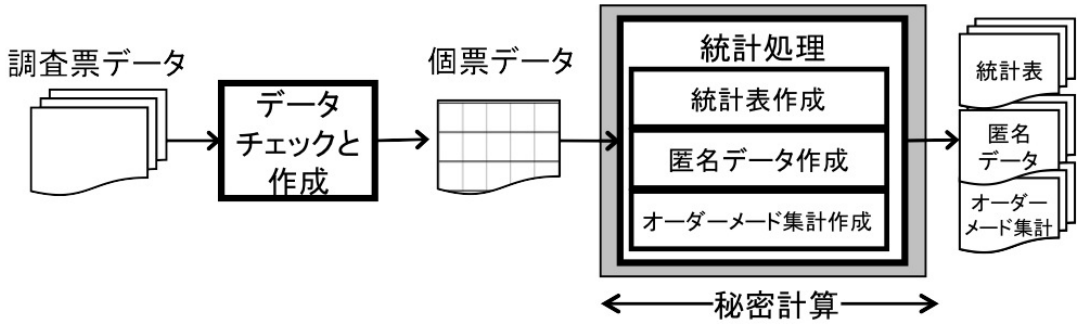
基本業務プロセスのはじめから終わりまで全ての処理に対して、秘密計算技術を用いて対応するケースが図９である。このモデルでは、調査票は記入された直後に暗号化され、処理の過程すべてが暗号化状態のままで行われ、最終的に定められた処理の結果のみが出力される。この完全なモデルでは処理結果以外の情報が、調査票データを作成した本人以外に漏れることがない。

図９　秘密計算を応用した統計の業務モデル１（調査票暗号化型）



次に、図 10 は、統計表等の作成業務にのみ秘密計算を適用するモデルである。本モデルは各種の統計処理や二次的利用の過程でのプライバシーを守ることができる。現在の公的統計の作成では、調査票のチェックから個票の作成の過程は安全性も含めて確立しているため、本モデル２（個票暗号化型）をベースに秘密計算技術の応用を検討することとした。

図 10　秘密計算を応用した統計の業務モデル２（個票暗号化型）



このモデルは、調査票データの回収からデータチェックと個票の作成に至るプロセスは安全に行われている前提で、二次的利用を推進するための秘密の保護及びデータを最小限にすることへの貢献が可能なモデルである。

5.2　適用モデルの検討

図 10 を元に、代表的な統計業務への適用を考察する。統計表作成、オーダーメード集計、「匿名化」データ作成に適用したモデルが図 11、図 12、図 13 である。

5.2.1　統計表作成モデル

図 11　統計表作成モデル

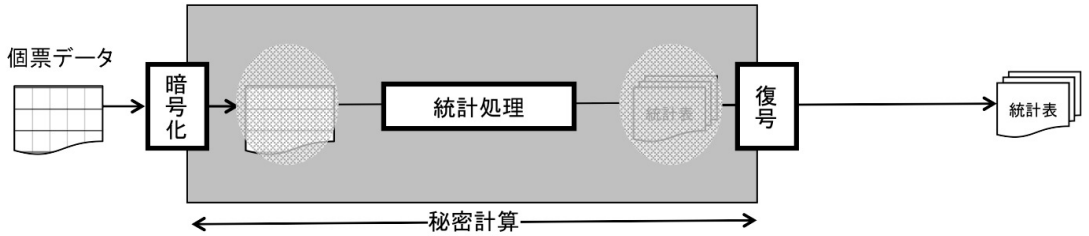


図 11 の統計表作成のモデルでは、個票データが暗号化されて秘密計算システムに与え

られ、秘密計算で統計表が作成され、その結果統計表だけが復号され出力される。このモデルでは個票の秘密を統計表を作成する一連のプロセスから秘匿することが利点である。本モデルを実現するためには、秘密計算システム用の統計表作成処理を開発する必要がある。

5.2.2 「匿名化」データ作成モデル

図 12 「匿名化」データ作成モデル

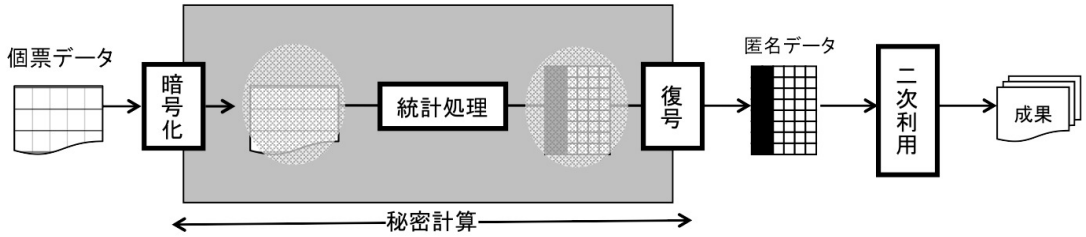
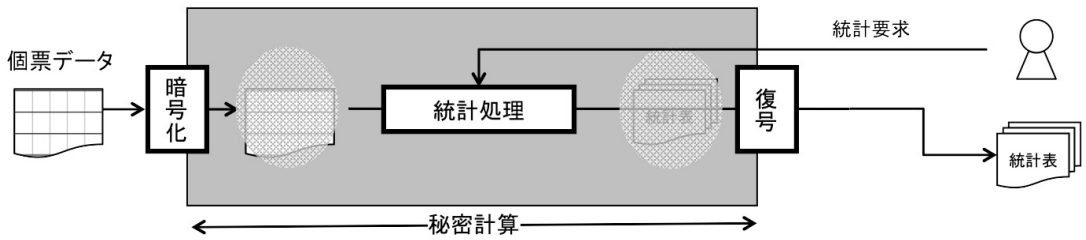


図 12 の「匿名化」データ作成のモデルは、統計表作成モデルと同様に個票データが暗号化されて秘密計算システムに与えられる。違いは秘密計算で行われる処理が、統計表作成ではなく「匿名化」データ作成であることである。このモデルでは個票の秘密を「匿名化」データを作成する一連のプロセスから秘匿することが利点である。本モデルを実現するためには、秘密計算システム用の「匿名化」データ作成処理を開発する必要がある。

このモデルは二次的利用のために開示する情報量が統計表作成モデルと比べて多い。秘密計算が担当する範囲を設計することで個票の秘密を守る程度が決まる。

5.2.3 オーダーメード集計モデル

図 13 オーダーメード集計モデル



次に図 13 に、オーダーメード集計に適用するモデルを示す。オーダーメード集計は、統計利用者のオーダーに基づいて集計を行うものであり、オーダーメード集計で統計表を

作成するのであれば、統計表作成モデル（図 11）に利用者からの統計要求を付け加えたものになる。課題は、統計表から秘匿すべき情報がもれることを防ぐための、作成される統計表の抑制（セル秘匿）や繰り返しの統計要求からもれる情報を抑制する仕組みの秘密計算上の実装である。

5.3 モデルに関する考察

個票データを暗号化して秘密計算を行うモデルをベースに、さらに秘密計算の役割分担の範囲を分類して比較した。

その結果、秘密計算の範囲が広いほどデータはより安全に保たれるが、一方、秘密計算の実装の難易度は上がり、さらに二次的利用者の自由度、例えば自分が好みの統計レポーティングソフトウェアを使いたいというニーズ、を満たすことはできなくなる。

また、秘密計算が実装すべき機能は、以下のように分類される。

表 3　秘密計算が実装すべき機能

1	登録クライアント	個票データの暗号化（秘密分散）
2	統計クライアント	統計要求と結果の表示（利用者側の入出力）
3	サーバ	一般的な統計処理の秘密計算実装
4	サーバ	統計表の作成の秘密計算実装
5	サーバ	「匿名化」データ作成処理の秘密計算実装
6	サーバ	「匿名化」データの秘匿の確認の秘密計算実装
7	サーバ	統計表からもれる情報の抑制の秘密計算実装

6　評価

秘密計算システムを統計研修所のミクロデータ共同利用研究室で構築し、評価を行った。

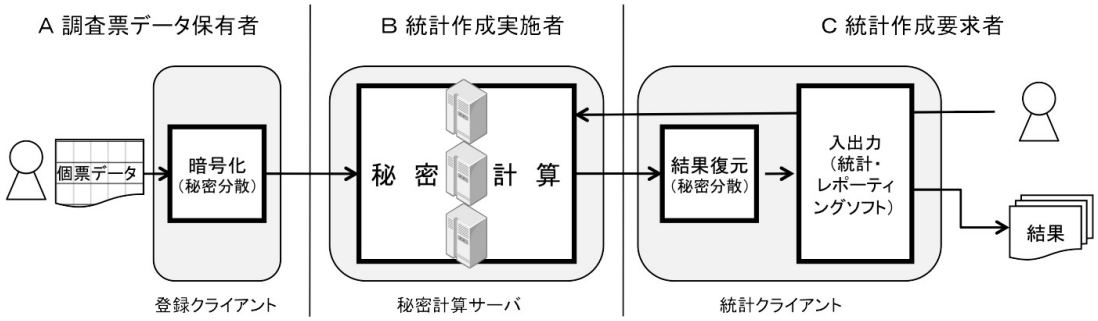
6.1　プロトタイプシステムのアーキテクチャ

図 14 に示す構成の秘密計算プロトタイプシステムを構築した。

登場人物の役割構成はオンデマンド集計の実現を意識して、A調査票データの保有者、B統計作成実施者、C統計作成要求者の３つに定義したうえで、Aが調査票を秘密分散してBに登録し、CがBに対して統計作成要求を行い、Bが秘密計算を行うモデルとした。

このモデルではAは調査票データをBに対して安全性を確保して提供でき、Bはデータの安全性を確保してCの統計作成要求に答えることができる。なお、このオーダーメード集計は、秘密計算システムで実装されている統計演算に限定してCがテーブル、属性、属性値などを指定して行うものである。

図 14　システムアーキテクチャ



6.2　プロトタイプシステムの構成

プロトタイプシステムの構築内容は以下の通りである。

登録クライアント

個票データを暗号化して登録するための秘密分散機能を構築した。本クライアントは、個票データのスキーマを指定して、スキーマに従ってセル単位で秘密分散を行う。5.3 節で定義した機能の 1 に相当する（表 4）。

秘密計算サーバ

登録され秘密分散された個票データに対して、統計処理を行う機能を持つ秘密計算サーバライブラリを構築した。5.3 節で定義した機能の 3 に相当する。4 から 7 に相当する統計表に関する複雑な機能は実装されていない。

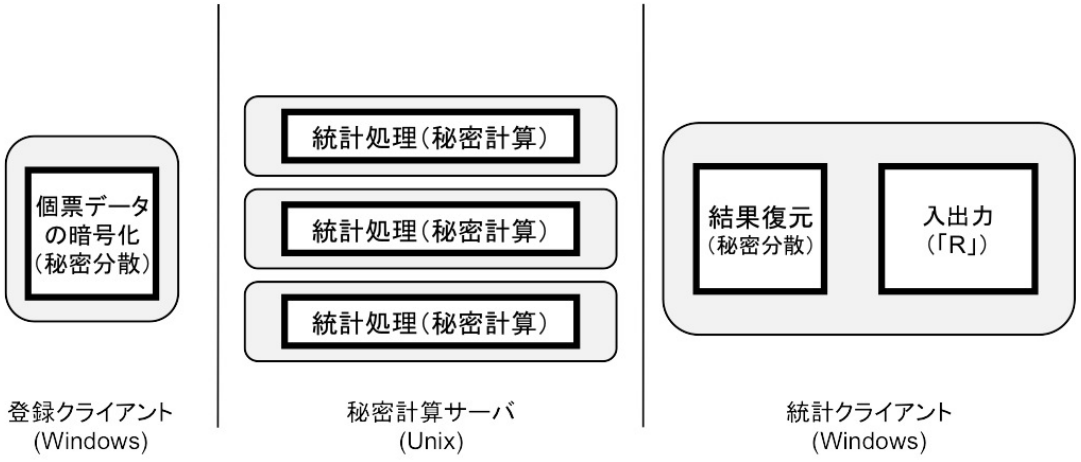
統計クライアント

統計作成要求者からのリクエストを受け付け秘密計算に計算要求を出す機能と、秘密計算からの暗号化された結果を受け取り、それを復元して表示する秘密分散の復元機能を構築した。5.3 節で定義した機能の 2 に相当する。

表 4　秘密計算プロトタイプシステムで実装した機能

実装	1	登録クライアント	個票データの暗号化（秘密分散）
実装	2	統計クライアント	統計要求と結果の表示（利用者側の入出力）
実装	3	サーバ	一般的な統計処理の秘密計算実装
未実装	4	サーバ	統計表の作成の秘密計算実装
未実装	5	サーバ	「匿名化」データ作成処理の秘密計算実装
未実装	6	サーバ	「匿名化」データの秘匿の確認の秘密計算実装
未実装	7	サーバ	統計表からもれる情報の抑制の秘密計算実装

図 15　秘密計算プロトタイプシステム



本秘密計算システムは、調査票のスキーマ及びレコード数公開型、すなわち登録されるデータ表の表側・表頭を秘匿せずにクライアントとサーバ間で共有するモデルを採用して