

## ＜資料編＞

### 資料 2-1 主催講演

「Synthetic Data（合成データ）の概要と  
調査業務・公的統計への応用」

（立正大学データサイエンス学部 教授 高部 勲）



# **Synthetic Data (合成データ) の概要と 調査業務・公的統計への応用**

**高部 勲**  
立正大学データサイエンス学部

1

## **Contents**

1. 自己紹介・本日の講演概要
2. 公的統計ミクロデータの概要
3. 合成データ (Synthetic Data) の概要
4. 合成データの作成の事例
5. 合成データの活用可能性

2

# 👉 1. 自己紹介・本日の講演概要

2. 公的統計ミクロデータの概要

3. 合成データ（Synthetic Data）の概要

4. 合成データの作成の事例

5. 合成データの活用可能性

3

## 自己紹介（DS学部ウェブサイト） [https://www.ris.ac.jp/ds/staff/009\\_isao-takabe.html](https://www.ris.ac.jp/ds/staff/009_isao-takabe.html)

立正大学データサイエンス学部  
Faculty of DATA SCIENCE



Isao Takabe

博士(統計科学)  
**高部 勲**  
たかべ いさお

専攻(担当分野)  
公的統計、統計科学

略歴

2001年

早稲田大学理工学部卒業

2008年

政策研究大学院大学政策研究科政策専攻修士課程修了、修士(公共政策)

2019年

総合研究大学院大学総合科学研究科統計科学専攻博士課程修了、博士(統計科学)

2002~2021年

総務省(統計局、大臣官房、統計研究研修所等)、独立行政法人統計センター

2020~2021年

統計数理研究所客員准教授

2020~2021年

滋賀大学特任准教授

2021年

現職

研究テーマ

(1)公的統計データ・企業データのデータリンケージ・統計的マッピング、(2)状態空間モデル・時系列同位モデル等に基づく消費関連データ(公的統計データ、POSデータ)の予測・経済指標の開発、(3)公的統計ミクロデータの高度な活用方法などについて研究しています。

担当科目、ゼミ

AⅠ入門Ⅰ、AⅠ入門Ⅱ、統計調査法、社会調査の設計と実施、社会調査実習Ⅰ、社会調査実習Ⅱ、ゼミナールⅠ、ゼミナールⅡ、ゼミナールⅢ、ゼミナールⅣ、卒業研究・卒業論文

企業データの統計的マッピング及び  
実業統計に関する研究

高部 勲

博士(統計科学)

東京大学大学院経済学研究科  
経済学専攻  
教授  
〒100-8302 東京都千代田区千代田1-8-8



ビジュアルでわかる  
**統計学のキホン**



統計学の“考え方”を  
スッキリ理解!

AI、IoT時代の企業、医療、マーケティング、経営戦略に  
さらに重要度を増す統計学。その概念の年々を今知る!

M&N

日本統計学会公式認定  
統計検定(統計調査Ⅱ)対応  
**経済統計の実際**

日本統計学会 編

Data Science  
日本統計学会公式認定  
統計検定(データサイエンス発展)対応  
**データサイエンス  
発展演習**

日本統計学会 編

4

(40)

## 総務省で統計指標の研究・開発を経験！

### データを提供する側に求められる視点と役割とは？

たかべ いさお  
高部 勲先生  
データサイエンス学部  
データサイエンス学科

#### Q 総務省ではどんな仕事をしていたの？

総務省では、主に統計局で、統計調査を実施する際の調査票・調査用品の作成や予算の要求、体制の整備などの調査の企画立案・実施に関する業務のほか、企業に関するデータベースの整備や、公的統計の作成・公表、新たな経済指標の開発など、「正確なデータを作る」ことに關する様々な業務に携わってきました。その過程で、時には、国会議員への説明やマスコミ対応、国会答弁を書くといった業務もしていました。

その中でも印象に残っているのが、先端的な統計学の知識と公的統計データを活用した新たな統計指標の研究・開発に関する業務で、当時組んだプログラムのコアな部分が今でも動いていて、都道府県別失業率や家計の消費支出に関する指数（CTI）などの結果が算出されています。

#### Q AIを導入する際に気を付けるべきことは？

業務などにAIを導入する際には、どのような機器やアルゴリズムを用いるか、といった技術面での検討も重要ですが、一方で、AIの学習の際に用いるデータそのものに関する検討もまた重要であり、データが持つ偏りや（データバイアス）、偏ったデータをAIに入力して読んだアルゴリズムを学習してしまう危険性（アルゴリズムバイアス）にも十分に配慮する必要があります。

例えばSNSのデータを扱う場合に、若年層が多いなど年齢層に偏りがある場合には、こうした点を考慮した調整が必要であり、また、企業の採用業務に関するAIを導入する場合には、過去の情報から特定の年齢や性別に偏った採用傾向を学習してしまわないように注意するなど、データの特性に応じた配慮が必要です。



#### 【担当科目】

- AI入門Ⅰ・Ⅱ
- 統計調査法
- 社会調査の設計と実査
- 社会調査実習Ⅰ・Ⅱ

#### 【専攻分野】

- 公的統計、経済統計
- 統計調査、社会調査
- EBPM（Evidence Based Policy Making: データに基づく政策立案）
- 統計科学、Data Science

#### 【その他】

- ✓ 総務省、内閣官房、和歌山県などの有識者委員・非常勤研究員等に從事
- ✓ 統計検定、専門社会調査士、ディープラーニングG検定などDS関連の資格多数取得

5

## 本日の講演概要

- 【前置き】公的統計・公的統計ミクロデータの概要
- 合成データ（Synthetic data）の概要・活用事例
- 公的統計ミクロデータなどを基にした合成データの試作・作成事例（中小企業データ・世帯データ）  
⇒本物のデータを一部欠損させて補完する手法
- 公的統計分野への活用可能性
  - 教育・研修用データ
    - ・データのイメージ把握
    - ・公的統計の学習用データ
  - 開発・検証用データ
    - ・アプリ開発のテスト用データ
    - ・集計プログラムのテスト用データ
    - ・欠測値補完の手法検証用のデータ

6

1. 自己紹介・本日の講演概要

## 👉 2. 公的統計ミクロデータの概要

3. 合成データ (Synthetic Data) の概要

4. 合成データの作成の事例

5. 合成データの活用可能性

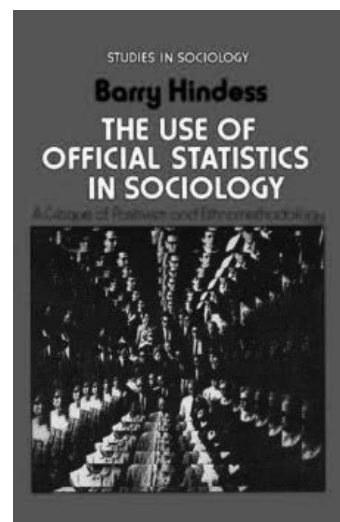
7

## 「公的」統計とは…

統計法 [平成十九年法律第五十三号]

第二条 (定義)

3 この法律において「公的統計」とは、行政機関、地方公共団体又は独立行政法人等が作成する統計をいう。



8

## 統計法の目的（第一条）

## 第一条（目的）

この法律は、公的統計が国民にとって合理的な意思決定を行うための基盤となる重要な情報であることにかんがみ、＜中略＞もって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する

※（参考）旧統計法〔昭和二十二年法律第十八号〕

(法の目的)

第一条 この法律は、統計の真実性を確保し、統計調査の重複を除き、統計の体系を整備し、及び統計制度の改善発達を図ることを目的とする。

9

**公的統計ミクロデータ（調査票情報）**

- 国の行政機関が実施した統計調査の結果について、調査対象の秘密の保護を図った上で、世帯単位や事業所単位といった集計する前の個票形式のデータ（マイクロデータ、調査票情報）を提供
- マイクロデータ（調査票情報）を用いることで、研究者の方々は、より自由で多様な分析を行うことが可能となるため、新たな発見につながることを期待（EBPMの観点からの利用等）



10

# 公的統計マイクロデータのオンライン利用

## マイクロデータの利活用なら **オンライン**

総務省など国の行政機関で実施した統計調査の結果は、ホームページ(政府統計の総合窓口 e-Stat)等を通じて広く一般の方にご利用いただいています。このような調査結果の提供に加え、公益性のある学術研究等にご活用いただくため、調査対象の秘密の保護を図った上で、世帯単位や事業所単位といった集計する前の個票形式のデータを提供しています。

この個票形式のデータを

マイクロデータ(調査票情報)と言います。

マイクロデータ(調査票情報)を用いることで、研究者の方々は、より自由で多様な分析を行うことが可能となるため、新たな発見につながることが期待されます。

マイクロデータ(調査票情報)のイメージ

Weight	Y_Income	L_Expendi	Food	Housing	LFW	Furniture	Clothes	Health	Transport	Education
895.2667	3917	201649	47756	16028	9652	6702	8088	726	21546	0
895.2667	6675	166381	34054	7416	26313	17062	6989	7637	20773	0
895.2667	6706	259736	84501	1927	10082	6741	5090	11015	53372	0
895.2667	2790	114511	41664	730	22358	5413	1205	5049	17411	0
895.2667	2577	193505	56981	3779	28747	4812	4243	751	16435	0
895.2667	3452	152109	34924	3418	8131	4164	6970	4247	47698	0
895.2667	3233	136900	49956	203	15429	3659	22843	4365	8684	0
895.2667	9252	192439	68882	2832	23042	2598	5714	2052	37006	0
895.2667	2359	138415	53591	753	13072	5140	1786	5416	11593	0
895.2667	2059	79179	32853	14134	7977	3017	2364	1607	7630	0
895.2667	2324	243835	60528	28118	16392	4652	10759	37060	17666	0
895.2667	4524	241539	104433	5253	40637	7711	13833	3853	43003	0
895.2667	4415	207854	95504	7687	13801	19702	7239	3939	9547	0
895.2667	4162	185110	59798	565	12146	14552	12301	2486	26726	0
895.2667	16647	219935	81572	3704	21164	4944	10282	6573	7103	0
977.1795	6760	176625	43112	5013	11780	1697	3008	4836	12295	848
977.1795	6614	130803	32336	3149	14723	21002	12075	3447	24273	140
977.1795	6849	395294	66755	9524	19962	19777	10598	7398	56370	80
977.1795	6813	284803	86655	13479	15121	22966	1134	3649	84535	551
977.1795	6595	260459	46614	1646	15044	6415	16207	3709	76074	69

(総務省統計データ利活用センター「オンライン利用のご案内」から引用)

([https://www.e-stat.go.jp/microdata/sites/default/files/share/data-use/onsite\\_intro.pdf](https://www.e-stat.go.jp/microdata/sites/default/files/share/data-use/onsite_intro.pdf))

※データは擬似データです

11

# 公的統計マイクロデータのオンライン利用

## 公的統計の マイクロデータを使って 分析してみませんか?

公的統計データの新たな利用方法

**オンライン**のご案内

公表されている集計結果からは  
分らないことも...

マイクロデータを使用  
分析によって...



公的統計の調査結果は、公表されている集計表に加え、  
マイクロデータを使用することにより、研究者の方々は  
更に多様な分析を行うことができるようになります。

### オンラインとは?

オンラインとは、情報セキュリティが確保された環境で、許可を受けた研究者がマイクロデータを用いて、独自の集計・分析を行うことができる専用室です。



入退室管理や監視カメラを備えたオンライン室

### オンラインでできること

- オンライン内に設置されたPCを用いて、利用申出を行った公的統計のマイクロデータを使った研究分析を行うことができます。
- また、研究者が用意したデータやプログラムも利用することができます。
- 分析した結果については、セキュリティ保護の観点等から、所定の審査を経た上で提供を受け、利用することができます。

(マイクロデータ利用ポータルサイト“miripo”から引用  
<https://www.e-stat.go.jp/microdata/>)

12



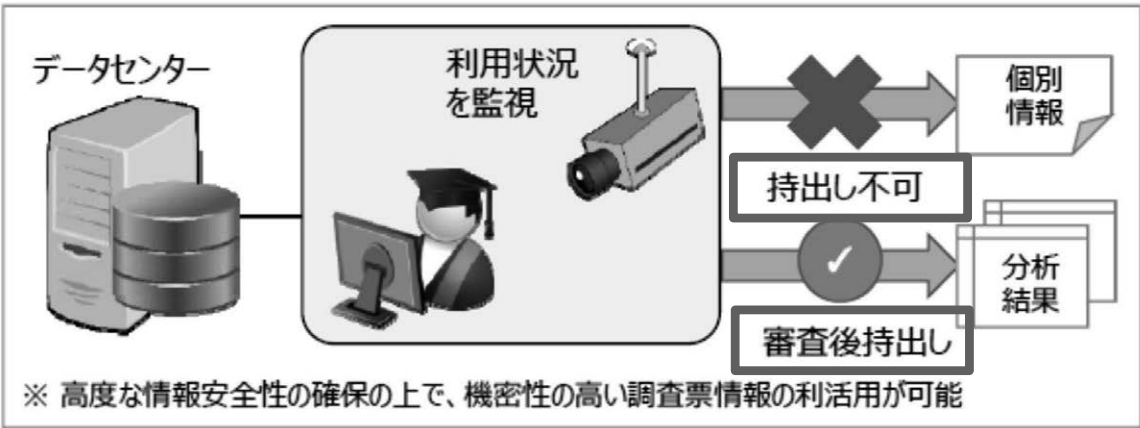
# 公的統計ミクロデータのオンライン利用

## 統計的マッチングやパネルデータの作成による 情報量の増加、利便性の向上

※有用なデータの作成には膨大な試行錯誤(探索)が必要

※分析結果の秘匿性(安全性)の確保が必要

⇒オンライン利用により探索的な分析と安全性の確保を実現



13

## 分析結果を持ち出す際の審査

表1 標準的なチェック内容

分析結果等の種類 <sup>*1</sup>		チェック内容 <sup>*2</sup>	申出者が提示する情報 <sup>*3</sup>
統計表	1. 度数表、度数の構成比表	①【度数】各セルが1以上10未満の調査客体から算出した値でないこと（加重なし <sup>*3</sup> ）	①【度数】各セルの度数（加重なし <sup>*3</sup> ）
		② <sup>*6</sup> 【度数】行計又は列計の90%超を占めるセルがないこと（加重なし <sup>*3</sup> ）	② <sup>*6</sup> 【度数】各セルの構成比（行計及び列計に占める割合）（加重なし <sup>*3</sup> ）
		③ <sup>*6</sup> 【度数】行計又は列計の90%超を占めるセルがないこと（加重あり <sup>*4</sup> ）	③ <sup>*6</sup> 【度数】各セルの構成比（行計及び列計に占める割合）（加重あり <sup>*4</sup> ）
	2. 1 数量表（総和）	①【度数】各セルが1以上10未満	①【度数】各セルの度数
	7. 線形回帰係数、非線形回帰係数	①【自由度】残差の自由度が10以上であること	①【自由度】残差の自由度
		②【データ】一つの調査客体のみに関するデータから作成していないこと（例：一つの調査客体のみに関する時系列データから作成していない）	②【データ】目的変数及び説明変数の説明資料

14

# 輸出入申告データのオンサイト利用

## 輸出入申告データを活用した共同研究の公募について（第4期）

いいね！ 6

✕ ポスト

令和6年9月2日

「経済財政運営と改革の基本方針2021」（令和3年6月18日閣議決定）等において、データの最大限の利活用を図ることとされていること等を踏まえ、財務省は、輸出入申告等に基づき税関が保有する輸出入申告データを、財務省の政策の検討に資するための学術研究に活用することとします。

このため、輸出入申告データを用いて、財務省の所掌に係る政策その他の内外経済に関するテーマに関する統計的研究を、財務総合政策研究所と共同して実施する研究者を公募いたします。

### 1. 募集テーマ

- (1) 国際貿易に関する研究（関税及び企業行動に関する分析を含む）
- (2) その他マクロ経済及び国際金融等に関する研究
- ※ (1)・(2) 合計で1～2件程度

### 2. 公募期間

令和6年9月2日（月）から同年12月2日（月）（必着）

※「個票データ等の利用申出意向表明書」については令和6年11月4日（月）（必着）

（輸出入申告データを活用した共同研究の公募について（第4期））

（財務省[https://www.mof.go.jp/about\\_mof/councils/kyoudou/index.html](https://www.mof.go.jp/about_mof/councils/kyoudou/index.html)）

15

# 輸出入申告データのオンサイト利用

## 問3. 個票データや研究の過程で作成された中間生成物を自身の所属する研究機関・研究室で利用できないでしょうか。

いいね！ 0

✕ ポスト

（答）

- 個票データは各輸出入申告別のデータであり、輸出入者の秘密の保護が強く求められることから、個票データ自体、およびそれを集計・分析する過程で生成された中間生成物については、財務総合政策研究所の施設内に設置された、財務総合政策研究所が提供する端末において利用していただきます。
- ただし、個票データを集計・分析する過程で生成された中間生成物のうち、ガイドライン別紙のチェック内容を満たしていることを財務省が確認したものは、「分析結果等」として、財務総合政策研究所の施設外に持ち出して利用することを可能としています。
- 分析結果等の利用にあたっては、不正アクセス等を防止するための措置を講じた上で、第三者に分析結果等の閲覧、利用、保管及び管理をさせないようにご注意ください。

（参照）

- ・ガイドライン第3の2

（輸出入申告データを活用した共同研究の公募について（第4期））

（財務省[https://www.mof.go.jp/about\\_mof/councils/kyoudou/index.html](https://www.mof.go.jp/about_mof/councils/kyoudou/index.html)）

16

# 匿名医療保険等関連情報データベース【NDB】のオンサイト利用

## 1. オンサイトリサーチセンターの拠点と利用要件

- オンサイトリサーチセンターとは、厚生労働省が指定する情報セキュリティ対策が講じられた施設において、厚生労働省が管理するNDBと通信回線で結ばれた端末の利用環境です。
- 全国に3拠点（厚生労働省、東京大学、京都大学）あります。
- 利用する場合、取扱者のうち一人以上は、以下の全ての要件を満たすことが必要です。

ア NDB第三者提供の特別抽出の利用経験を有する者  
イ SQLの知識を持ち、自身でデータベースを操作可能な者

「第44回レセプト情報等の提供に関する有識者会議 資料（オンサイトリサーチセンター（厚生労働省）の今後の方針について）」抜粋

担当者及び取扱者が上記要件を満たすことを示すために、様式1の「(4)-6 研究概要」にア、イの要件に関する具体的な経験等を記す必要があります。



（【NDB】匿名医療保険等関連情報データベースの利用に関するホームページ「NDBの利用を検討している方へのマニュアル」）  
([https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou\\_iryoku/iryohoken/reseputo/index.html](https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/iryohoken/reseputo/index.html))

17

# 匿名医療保険等関連情報データベース【NDB】のオンサイト利用

## 3. 利用にあたっての制約事項

### □匿名レセプト情報の参照

匿名レセプト情報については、NDBに格納されているデータのうち、直近10年分のみ参照可能となります。

### □情報の持ち出し

- ・ NDBデータについて、原則外部への持ち出しは禁止されております。
- ・ オンサイトリサーチセンターからの中間生成物又は最終生成物を含めた情報を持ち出す際には、事前に任意の様式で厚生労働省へ報告することとし、厚生労働省は、当該研究の持ち出し予定情報とあらかじめ承諾された形式が整合的であるか確認します。また、必要に応じて専門委員会の委員が確認を行います。（詳細はガイドラインを参照ください）

＜公表予定物を持ち出す場合の報告内容（原則申請時に記載）＞

- ✓ 使用したデータ、解析内容・手法、公表予定物のひな形

＜中間生成物を持ち出す場合の報告内容（原則申請時に記載）＞

- ✓ 使用したデータ、解析内容・手法（今後予定している解析内容・手法含む）、中間生成物のひな形（抽出条件・集計表のひな形等）、中間生成物の形式（個票・集計表）に応じたセキュリティ環境等

### □オンサイトリサーチセンター内での作業の外部委託の禁止

オンサイトリサーチセンター内での作業については外部委託することは認められません。

### □同期間に第三者提供と同一担当者による申出は禁止

第三者提供にて提供を受けたNDBデータの利用期間に、同一担当者によるオンサイトリサーチセンターの利用依頼申出は「重複提供」に当たるため、認められません。

※同様に、オンサイトリサーチセンターに利用申出を実施している担当者は、第三者提供の申出はできません。

（【NDB】匿名医療保険等関連情報データベースの利用に関するホームページ「NDBの利用を検討している方へのマニュアル」）  
([https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou\\_iryoku/iryohoken/reseputo/index.html](https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/iryohoken/reseputo/index.html))

18

1. 自己紹介・本日の講演概要

2. 公的統計マイクロデータの概要

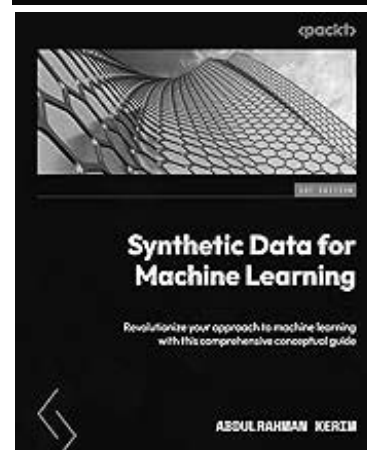
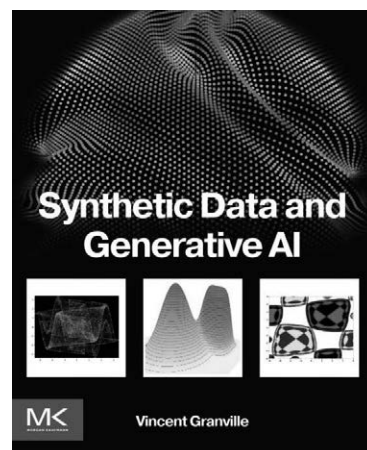
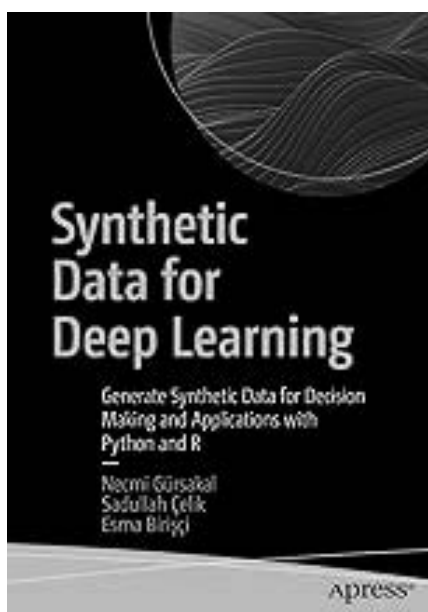
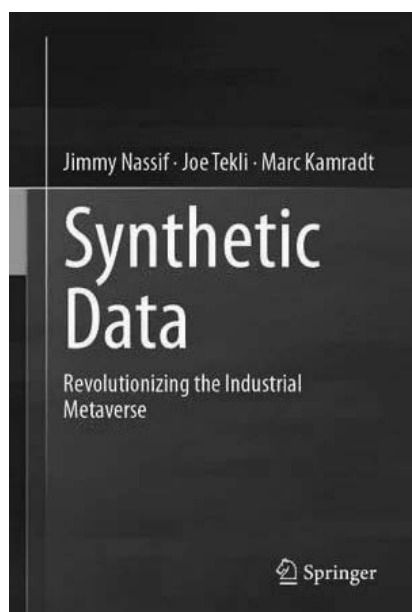
👉 3. 合成データ (Synthetic Data) の概要

4. 合成データの作成の事例

5. 合成データの活用可能性

19

## 合成データとは



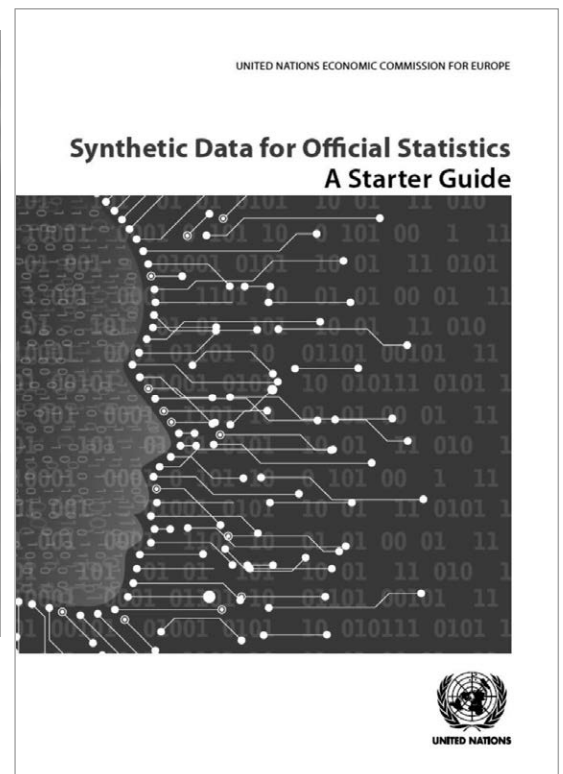
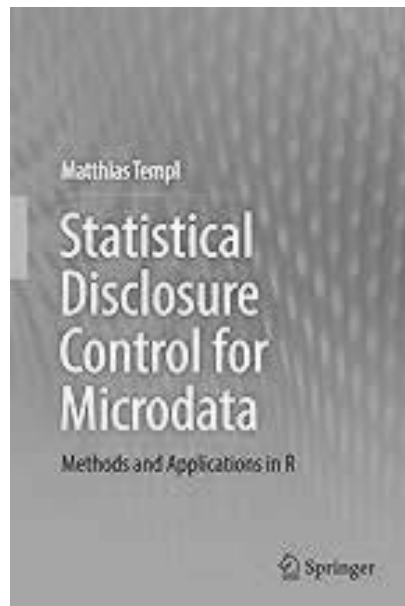
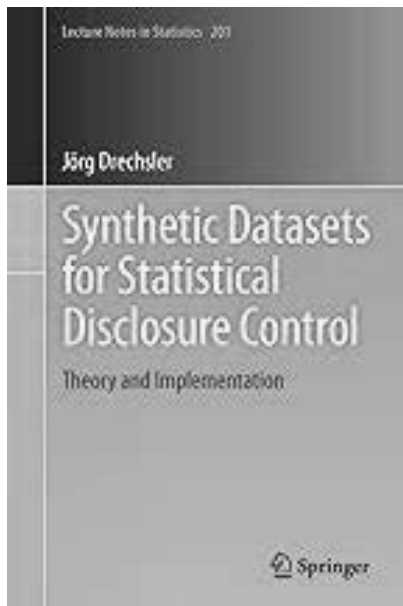
20

# 合成データとは

- 機械学習には膨大なデータが必要  
⇒膨大なデータを準備するのはコスト・時間がかかる  
⇒セキュリティ・プライバシー上の問題もある
- 大学の研究機関やAI開発企業が研究用途に公開している画像・テキストデータ  
⇒データの偏り（バイアス）などの問題
- これらの課題に対処するために「合成データ（Synthetic Data）」が用いられる  
⇒本物のデータを模して作成された人工的・擬似的なデータ

21

# 合成データとは



22

# 合成データとは

## synthpop: Bespoke Creation of Synthetic Data in R

Beata Nowok  
University of Edinburgh

Gillian M Raab  
University of Edinburgh

Chris Dibben  
University of Edinburgh

### Abstract

In many contexts, confidentiality constraints severely restrict access to unique and valuable microdata. Synthetic data which mimic the original observed data and preserve the relationships between variables but do not contain any disclosive records are one possible solution to this problem. The **synthpop** package for R, introduced in this paper, provides routines to generate synthetic versions of original data sets. We describe the methodology and its consequences for the data characteristics. We illustrate the package features using a survey data example.

**Keywords:** synthetic data, disclosure control, CART, R, UK Longitudinal Studies.

This introduction to the R package **synthpop** is a slightly amended version of Nowok B, Raab GM, Dibben C (2016). **synthpop: Bespoke Creation of Synthetic Data in R**. *Journal of Statistical Software*, **74**(11), 1-26. doi:10.18637/jss.v074.i11. URL <https://www.jstatsoft.org/article/view/v074i11>.

23

# 合成データとは

経済学論叢 (中央大学) 第64巻第3・4合併号 (2024年1月)

147

### わが国の公的統計における 合成データの展開可能性に関する一考察 ——事業所・企業系の統計調査を例に——

伊藤 伸 介\*  
横溝 秀 始\*\*

1. はじめに
2. わが国における公的統計の合成データの作成可能性について
3. ミクロアグリゲーションの特徴について
4. 合成データに対する匿名性と有用性の評価指標について
5. 合成データの作成方法の有用性の検証——経済センサス活動調査を用いて
6. むすびにかえて

#### 1. はじめに

海外では、近年合成データ (synthetic data) の作成とその実用化に向けた研究に対する関心が高まっている<sup>1)</sup>。合成データは、元になるデータからその分布特性が近似するように属性値を新たに生成することによって作成され、個人情報保護が確保されたミクロレベルの擬似的なデータと位置付けられる (Templ (2017, p.157))。合成データは、個票データに含まれる個人情報の特定につながりうる変数・レコード群に対して、機密的手法を含む各種の匿名加工の手法を施して作成される匿名化ミクロデータ (anonymized microdata) とは異なる。合成データの作成に関しては、統計モデルに基づくパラメトリックな手法の適用だけでなく、CART (Classification And Regression Tree) (Breiman et al. (1984)) 等によるノンパラメトリックな手法も適用されてきた (Re-

統計研究叢報 第79号 2022年3月 (111-130)

### 合成データの考え方にに基づく公的統計疑似ミクロデータの 作成方法の検討

高部 勲†

A study of creating synthetic data that contributes to promoting the utilization of official statistics microdata

TAKABE Isao

公的統計ミクロデータの更なる利用促進のために、教育・訓練用の疑似的なミクロデータの必要性が指摘されている。このような目的で現状、提供されている一般用ミクロデータは、元のデータの持つ変数間の相関などの構造を可能な限り保持するという点で課題がある。そこで、諸外国で研究が行われている合成データに関するモデルベースの手法を基に、現行の法令・制度上の制約を満たしつつ元のミクロデータの構造を可能な限り保持した疑似的なミクロデータを作成する方法について、分析・検討を行った。その結果、元のデータの匿名性に配慮した上で、中間的な集計結果や回帰分析などの結果を事前に公開することにより、現行の制度に沿った形で、元のデータの持つ構造を保持した疑似的なデータを作成・提供できる可能性があることが示された。

キーワード 公的統計、ミクロデータ、合成データ

24

# 合成データとは

- 合成データは、元になるデータからその分布特性が近似するように属性値を新たに生成することによって作成され、個人情報秘匿性が確保されたミクロレベルの擬似的なデータと位置付けられる（Templ（2017, p.157））。
- 海外では、公的統計を対象にした合成データの方法論に関する様々な適用事例が存在する。
- 欧州統計局（Eurostat）：EU-SILCにおいて、統計的なモデルに基づくシミュレーションによる合成データの生成技法を用いて一般公開型ファイル（Public Use File=PUF）を作成・公開している（伊藤（2018））。
- Rの合成データ作成用のパッケージであるsynthpop（Nowok et al.（2016））の提供

経済学論纂（中央大学）第64巻第3・4合併号（2024年1月）

「わが国の公的統計における合成データの展開可能性に関する一考察—事業所・企業系の統計調査を例に」伊藤伸介・横溝秀始

25

# 合成データとは

 あなたの中心にある。  
一般社団法人 日本マーケティング・リサーチ協会

フリーワード検索 🔍 会員ログイン 会員登録 ENGLISH

JMRAのご紹介 JMRA会員について 入会のご案内 綱領・ガイドライン 委員会活動 事業活動

一般社団法人 日本マーケティング・リサーチ協会 > 事業活動 > 業界・海外動向 > 海外の最新動向 > 合成データ（Synthetic Data）の現在地と今後の課題

事業活動

セミナー

カンファレンス

イベント

説明会

刊行物等（メールマガジン）

業界・海外動向

> 経営業務実態調査

> 国内の最新動向

合成データ（Synthetic Data）の現在地と今後の課題

合成データ（Synthetic Data）の現在地と今後の課題

ESOMAR GMR日本アンバサダー ノノ瀬 裕幸

<https://www.jmra-net.or.jp/activities/trend/international/20250121.html>

26

# 合成データとは

## 1. プライバシー／機密保護

合成データは、実際の個人情報を含まないため、データ漏洩のリスクを減らし、個人情報保護法を含む各種規制に対応しやすくなります。特に医療系や金融系など、機微な情報や機密情報が重要な分野で役立つとみられています。

## 2. データ収集コストの削減

調査テーマや目的によっては、現実のデータを収集することに多大な時間や費用がかかる場合があります。代替手段として合成データを使用すれば、必要なデータを素早く、安価に生成・利用することができます。

## 3. モデルのテストとトレーニング

これは生成AIの普及に伴って活用度の増した分野です。合成データはAIや機械学習のモデルを訓練するためにも使われています。例えば、自動運転車のシステムをテストする際に、現実の事故データを使う代わりに合成データで様々なシナリオを再現し、シミュレーションを行うことができます。そのほかにも、自然科学分野の各種実験領域では盛んに活用されるようになってきているようです。

<https://www.jmra-net.or.jp/activities/trend/international/20250121.html>

27

# 合成データの重要性

## ① プログラムテスト用データとしての活用

- 公的統計ミクロデータの利用申請を行う前にデータの利活用のイメージを把握（※オンサイト利用）
- 各種のアプリケーションの開発・テスト
- 匿名加工情報等の作成におけるニーズの事前把握 など

## ② データを利用した教育への活用

- 経済・社会の計量モデルは、個人・企業等を単位としているものが多い（重回帰分析など）  
⇒ 集計データではないレコード単位のデータの重要性  
⇒ 元データと結果がそれほど変わらないのが望ましい
- 複雑な手続きを必要とせず自由に利用できるデータの重要性

28



# 疑似データの重要性（教育への活用）

## 数理・データサイエンス・AI リテラシーレベルの教育方法

➤「導入」「基礎」「心得」「選択」のそれぞれの分類ごとに、推奨される具体的な教育方法を以下のとおりまとめた。

導入	<b>1. 社会におけるデータ・AI利活用</b> ・データ・AI利活用事例を紹介した動画（MOOC等）を使った反転学習を取り入れ、講義ではデータ・AI活用領域の広がりや、技術概要の解説を行うことが望ましい ・学生がデータ・AI利活用事例を調査し発表するグループワーク等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4
基礎	<b>2. データリテラシー</b> ・各大学・高専の特徴に応じて適切なテーマを設定し、 <u>実データ（あるいは模擬データ）を用いた講義を行うことが望ましい</u> ・実際に手を動かしてデータを可視化する等、学生自身がデータ利活用プロセスの一部を体験できることが望ましい ・必要に応じてデータハンドリングスキルを埋めるためのフォローアップ講義（補講等）を準備することが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4
心得	<b>3. データ・AI利活用における留意事項</b> ・身近で起こったデータ・AI活用における負の事例を通して、データ駆動型社会のリスクを自分ごととして考えさせることが望ましい（必要に応じてMOOC等の活用も検討する） ・データ・AIが引き起こす課題についてグループディスカッション等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4
選択	<b>4. オプション</b> ・本内容はオプション扱いとし、大学・高専の特徴に応じて学修内容を選択する ・各大学・高専の特徴に応じて適切なテーマを設定し、 <u>実データ（あるいは模擬データ）を用いた講義を行うことが望ましい</u> ・学生が希望すれば本内容を受講できるようにしておくことが望ましい（大学間連携等）	教育方法（例）※ 1, 2, 3, 4
		1, 2, 3, 4

※上記の「教育方法」欄の手法・形式 1～4 については次頁以降を参照

19

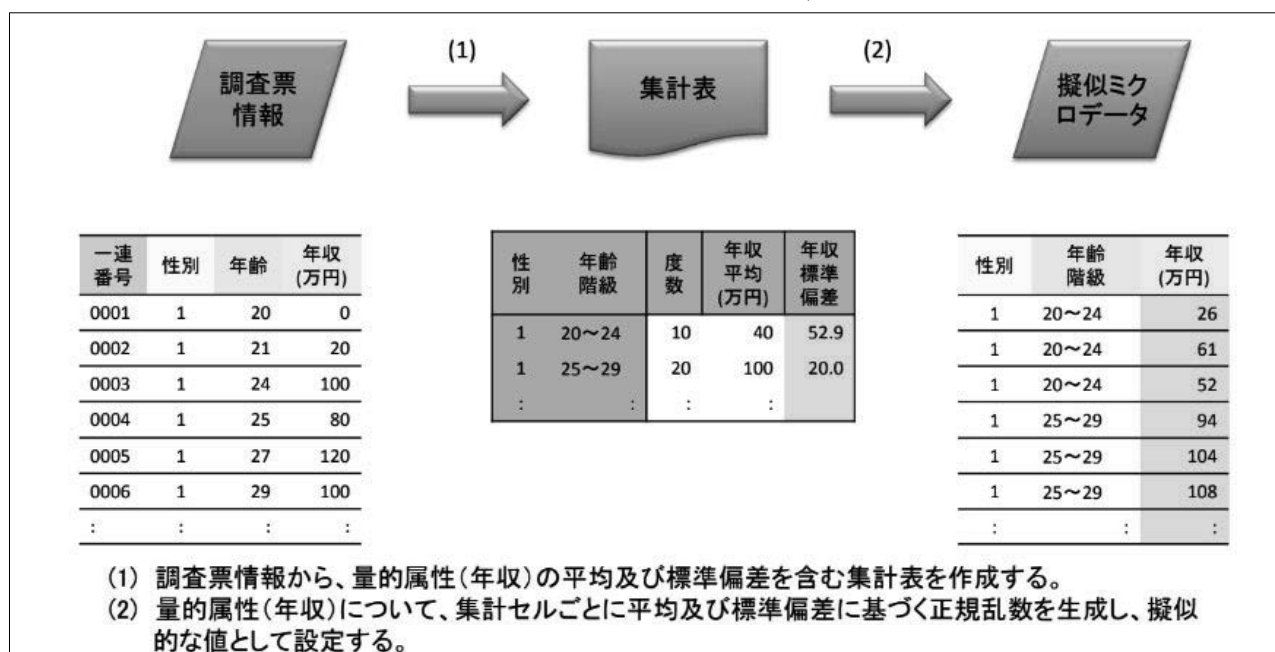
「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム」  
（2020年4月、数理・データサイエンス教育強化拠点コンソーシアム） から引用

29

## 一般用マイクロデータ

（「一般用マイクロデータ（仮称）の作成及び利活用について」  
（2015年度統計関連学会連合大会）から引用）

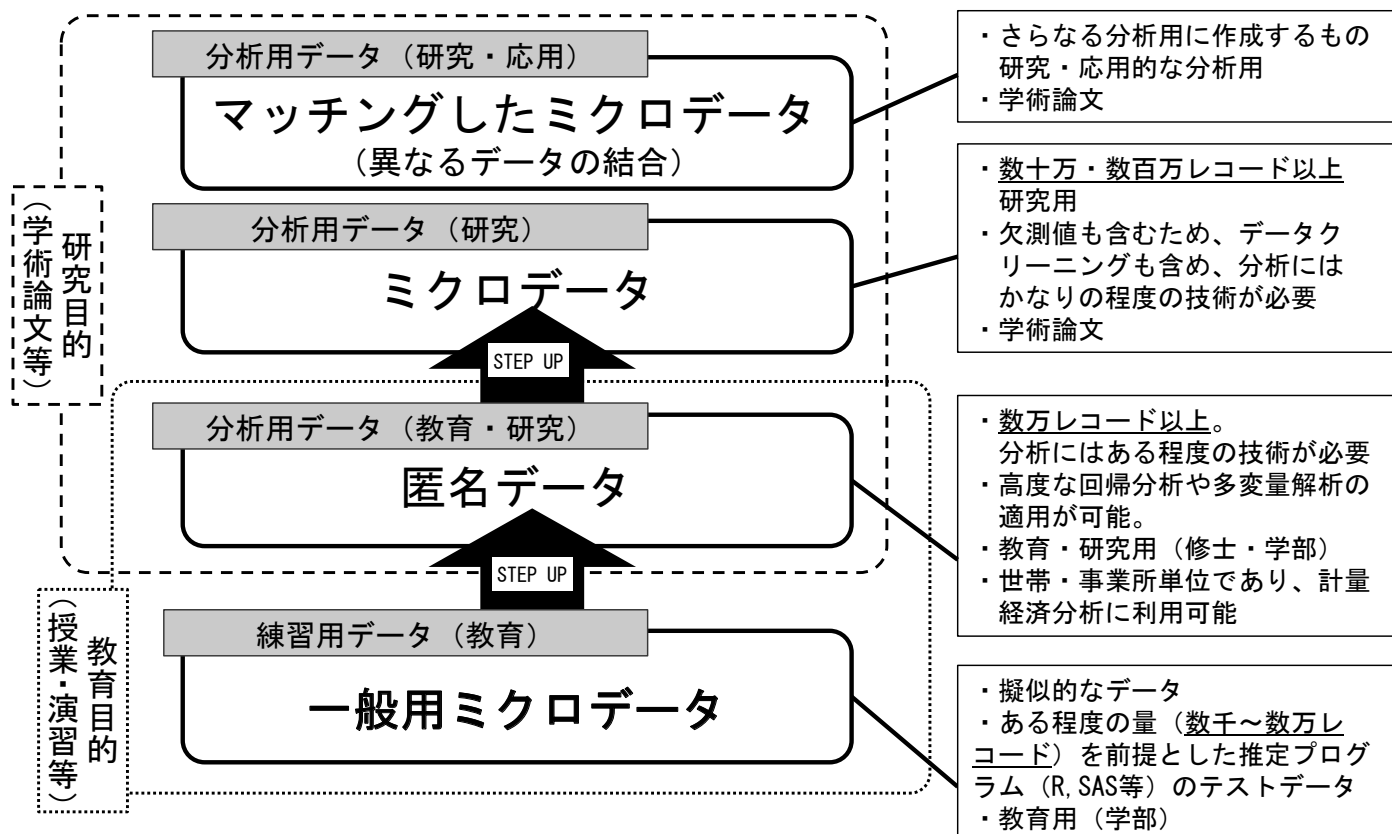
- 現行の制度上、マイクロデータから直接的にレコード単位の疑似データを作成・提供することはできない
- 中間的な集計表を公表し、そこから疑似データを作成



⇒ 変数間（離散・連続）の関係（回帰モデル）  
を考慮した疑似データの作成が課題

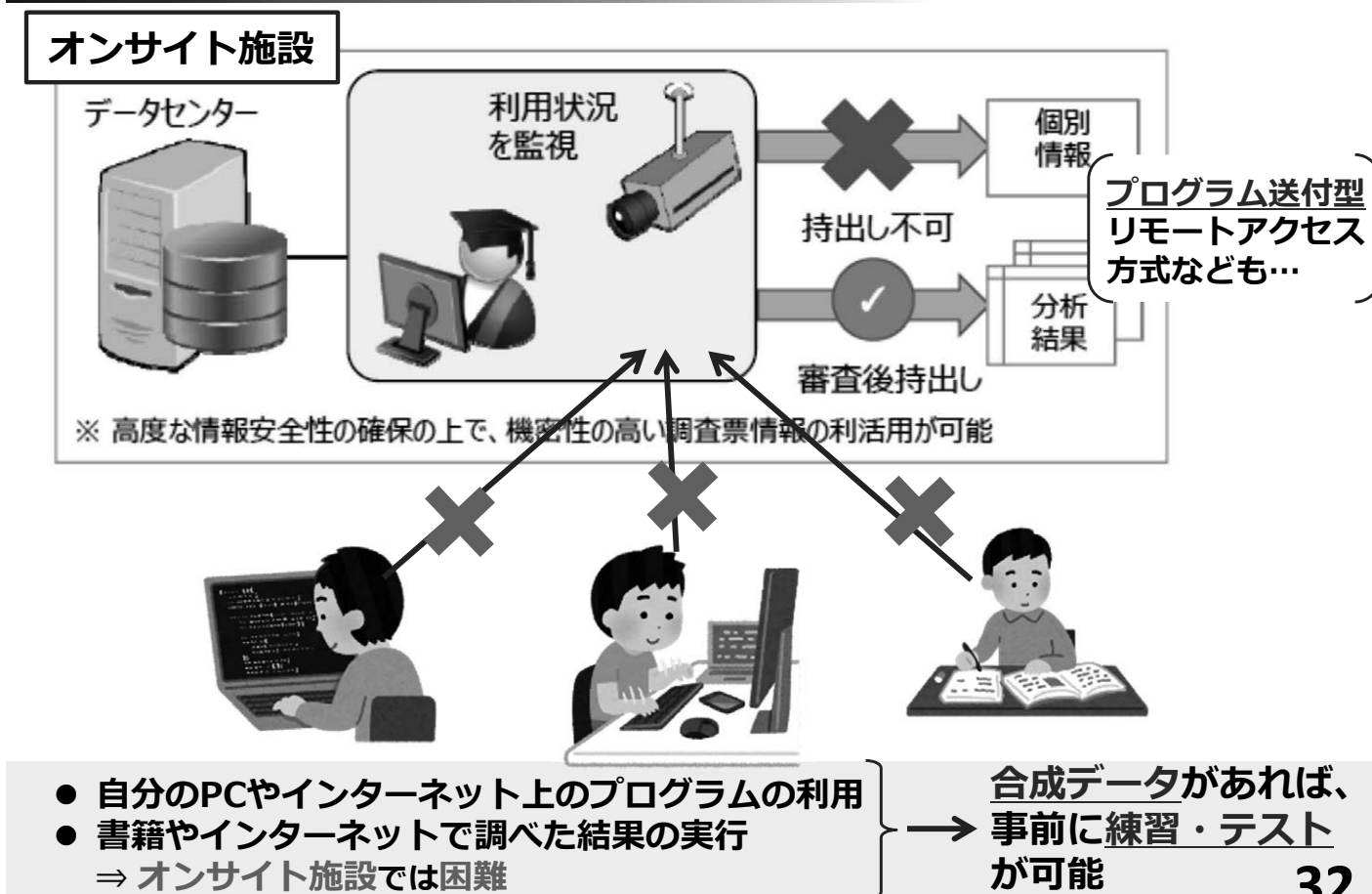
30

# 公的統計マイクロデータの概要



31

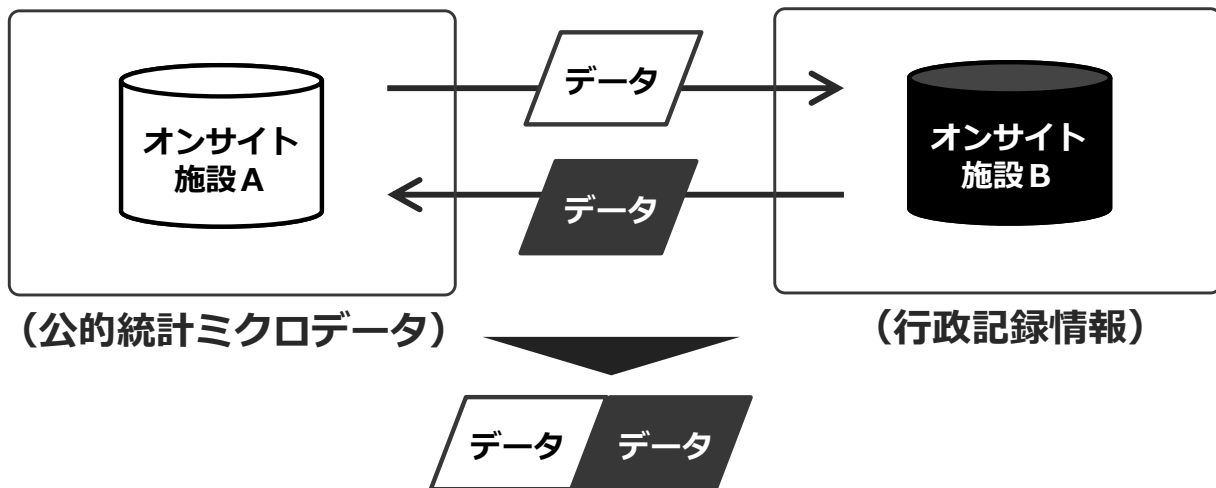
## 合成データの重要性（プログラムテスト用）



32

# オンサイト利用におけるデータ共有の課題

- 統計的マッチングが想定している状況（データの共有）



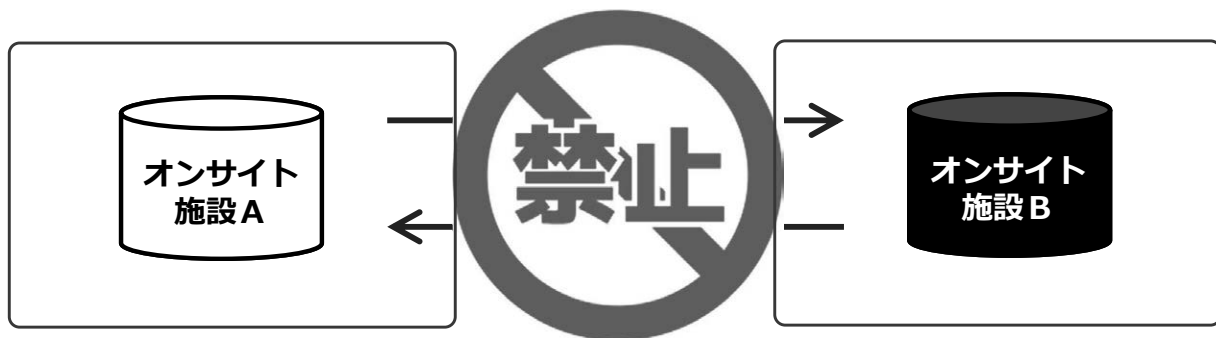
- 結合したデータを用いてモデルを推定・分析

- ✓ 世帯の様々な情報を用いた経済・社会・医療に関する分析
- ✓ 企業の生産性に関する情報と財務諸表に関する情報を組み合わせた分析
- ✓ アンケートデータとPOSデータ（レシート情報）を組み合わせたマーケティングに関する分析

33

## データ共有の課題

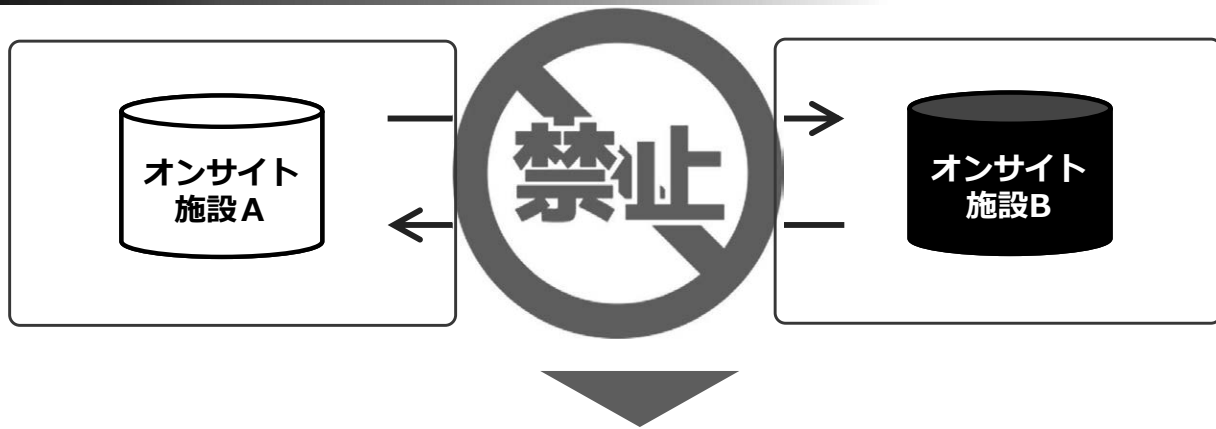
- 統計的マッチングが想定している状況（データの共有）



- 各オンサイトでは、法令・指針等により、データそのものの持出し・共有は不可  
⇒各オンサイトの全てのデータ・変数を用いたデータの結合や分析は不可能

34

# オンサイト利用におけるデータ共有の課題



## 元のデータの構造をある程度保持した疑似データの共有



35

# 疑似データの重要性（ニーズ把握）

## 行政のデジタル化推進

### 現状・課題

- 新型コロナウイルス感染症対応において、行政のデジタル化の遅れが顕在化
- 様々なニーズや目的に応じて社会のデータ利活用を推進することが不可欠

### 行政手続オンライン化の環境整備

#### ■ マイナンバーカードの普及促進

出張申請受付などの実施等を通じデジタル社会の基盤となるマイナンバーカードの普及を加速

#### ■ デジタル化有識者会議

県内市町村が一体的にデジタル化に取り組む方向性を明示

#### ■ 行政手続オンライン化の推進

国全体のデジタル化の方針を踏まえながら、県内市町村と一体となり、先行的に整備が可能な業務からオンライン化を一気に進める

**支援策** 県内デジタル行政専門人材派遣制度 **新規**  
・ 県内協力企業の I C T 専門人材を技術アドバイザーとして派遣

様々な住民サービスの手続がオンラインで完結することにより  
**便利で豊かな県民生活の実現**

### 【今後のロードマップ（想定）】

- 2021 年度**
- 先行的に整備が可能な申請手続のオンライン化
- 2022 年度**
- マイナンバーカードを用いる手続のオンライン化  
(例) 子育て関係(15手続)、介護関係(11手続) など
  - 全県(国)民によるマイナンバーカード取得

## (和歌山県「令和3年度新政策」から引用)

### データ利活用推進

#### ■ 先進的な分析・研究の促進（産学官で連携）

【例】空き家対策のためのデータ利活用モデルの構築  
(国、県、和歌山市、東京大学の共同研究)

#### ■ 庁内データ共有化モデルの構築 **≧全国初≧**

全国に先駆け行政データを匿名化し、二次的利用を図る部局横断的な利活用モデルを構築するとともに、他自治体への普及を促進

#### ■ 行政データ提供スキームの構築 **≧全国初≧**

行政データの民間活用を推進するため、匿名化技術等を活用し、分析ニーズに応じた加工データを提供

##### ① サンプルデータ

データレイアウトのみが  
そのまま、ランダムに  
生成した架空のデータ

広く一般に公開

データレイアウトの確認

##### ② 疑似データ **新規**

元データの特徴(分布等)  
を保持させながら生成した  
架空のデータ

有料

データの有用性の確認

##### ③ 非識別加工情報

特定の個人の識別が  
できないよう加工した  
実際のデータ

有料（審査有）

正式な分析・研究に活用

民間企業による新商品開発や販路開拓により  
**県内産業の活性化、地域課題の解決を実現**

36

# 参考：データ利活用に係る地方公共団体表彰

【特別賞】（第5回（2020年度））

茨城県  
つくば市

「疑似市民データ」を活用したアイデアソンの開催

- ✓ 街づくりのアイデアや地域課題の解決法を広く募るアイデアソンのイベントを開催
- ✓ 課題解決の検討には住民データを参考に作成した疑似データなどを活用

## 地域課題解決のアイデアソン開催

市民が自ら課題解決を考えていく、ディスカッションと解決策提案の機会を提供することを目的として開催

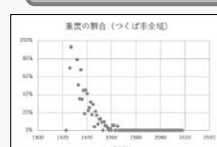
- ・2019年度テーマ：高齢者福祉に関わる課題解決
- ・市民、企業、大学、行政関係者など、延べ86名が参加
- ・GISやExcelを使った疑似データの可視化、相関・回帰分析の手法等によりデータ活用を検討



- ・アイデアソンでは、日常生活圏域（中学校区）ごとにクラスター分析することに着眼し、非識別加工情報で推察される匿名加工よりも非常に強いノイズやスワップを入れる加工を施し、疑似データ（非個人情報）を作成
- ・参酌データ（元データ）とは全く違うデータだが、クラスターごとの集計や比率は同じになるように、専門家の意見を取り入れて、ノイズ加工を施した。

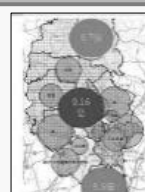
自治体の中でしか検討することができない詳細なデータに基づくディスカッションが可能に

## 地域貢献につながるアイデアの創出

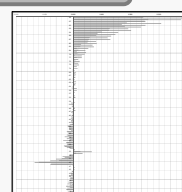


<年齢と重度介護者割合の相関>

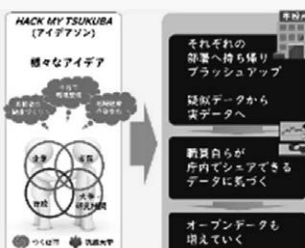
※疑似データのため結果は現状に沿っていません。



<65歳人口/10歳以下人口を可視化>



<行政区別年齢別人口データからコーホート分析>



<挙げられたアイデア例>

- ・ひとり暮らし高齢者の困りごと解決
- ・多世代コミュニケーションの場
- ・助けられる人と助ける人のマッチング
- ・動ける高齢者の活躍

<期待される効果>

自治体だけでは思いつかないアイデアを市職員が持ち帰り政策立案にいかす循環が生まれる強固な体制を構築

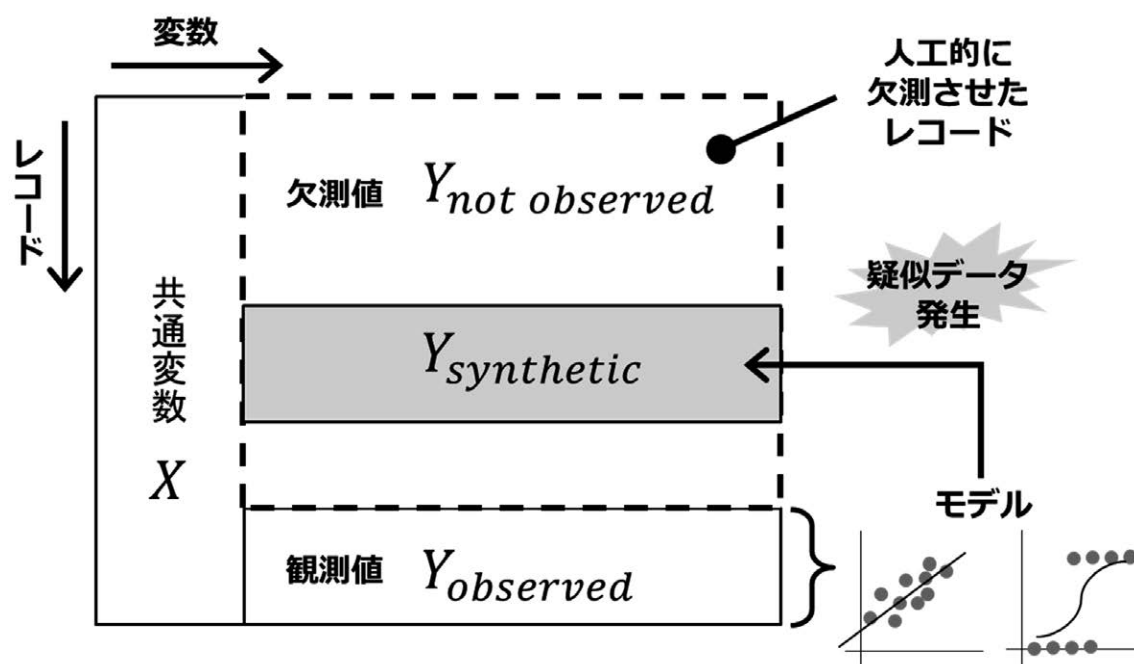
<活用データ> 国勢調査、国土数値、つくば市オープンデータ、住民基本台帳データ、レセプトデータ、健診結果情報 等

総務省報道発表資料（令和2年10月19日）から引用

37

## Synthetic Data

- 一部のレコードを人工的に欠測させ回帰モデル、ロジットモデル等により逐次的に疑似データを作成  
⇒変数間の関係を保持したデータの作成が可能  
（Rのパッケージも提供されている（“synthpop”））



38

1. 自己紹介・本日の講演概要

2. 公的統計マイクロデータの概要

3. 合成データ (Synthetic Data) の概要

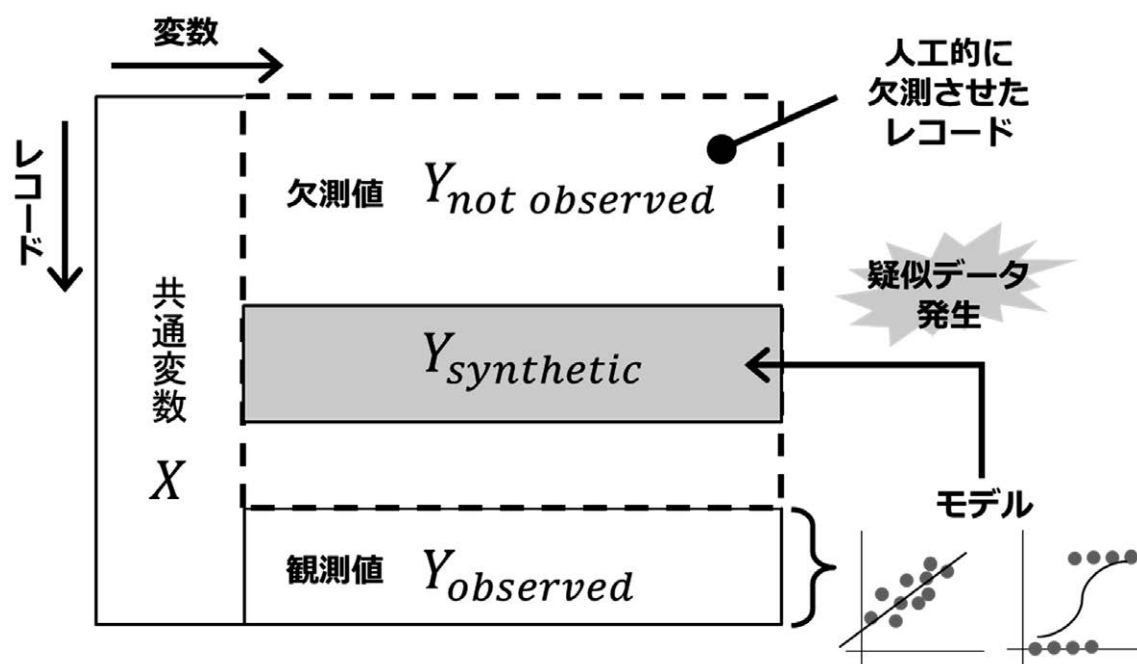
👉 4. 合成データの作成の事例

5. 合成データの活用可能性

39

## Synthetic Data 【再掲】

- 一部のレコードを人工的に欠測させ回帰モデル、ロジットモデル等により逐次的に疑似データを作成  
⇒ 変数間の関係を保持したデータの作成が可能  
(Rのパッケージも提供されている (“synthpop”))



40

# 疑似データの作成方法

- ただし現行制度上、マイクロデータから直接的にレコード単位の疑似データを作成・提供できない  
⇒ 現行の制度に合わせた形でSynthetic Dataを作成するための工夫が必要
- 本研究では、事前に集計・推定した統計表とモデルの結果を基に、Synthetic Dataの方法に基づいて疑似データを作成する方法について検討  
⇒ どのような集計結果、モデルの結果を公表すれば、秘匿性を確保した上で、元データの構造を保持した疑似的なデータが作成できるかを検討

41

## 合成データの試作①（中小企業データ）

### ➤ 商用データ(平成24年2月分)

- ある県のデータを使用
- サイズ：7,558レコード
- 連続変数：
  - ・従業員数（対数変換）
  - ・売上高（対数変換）（※資本金は使用しない（特定の値に集中、後述））
- 離散変数：
  - ・産業（建設業、製造業、小売業、それ以外）
  - ・地域（3地域）
  - ・経営組織（株式会社、有限会社）
  - ・開設年（～1984年, 1985年～1994年, 1995年～）

42

# データ作成の流れ

## ● 周辺分布（地域(3)×産業(4)）の集計

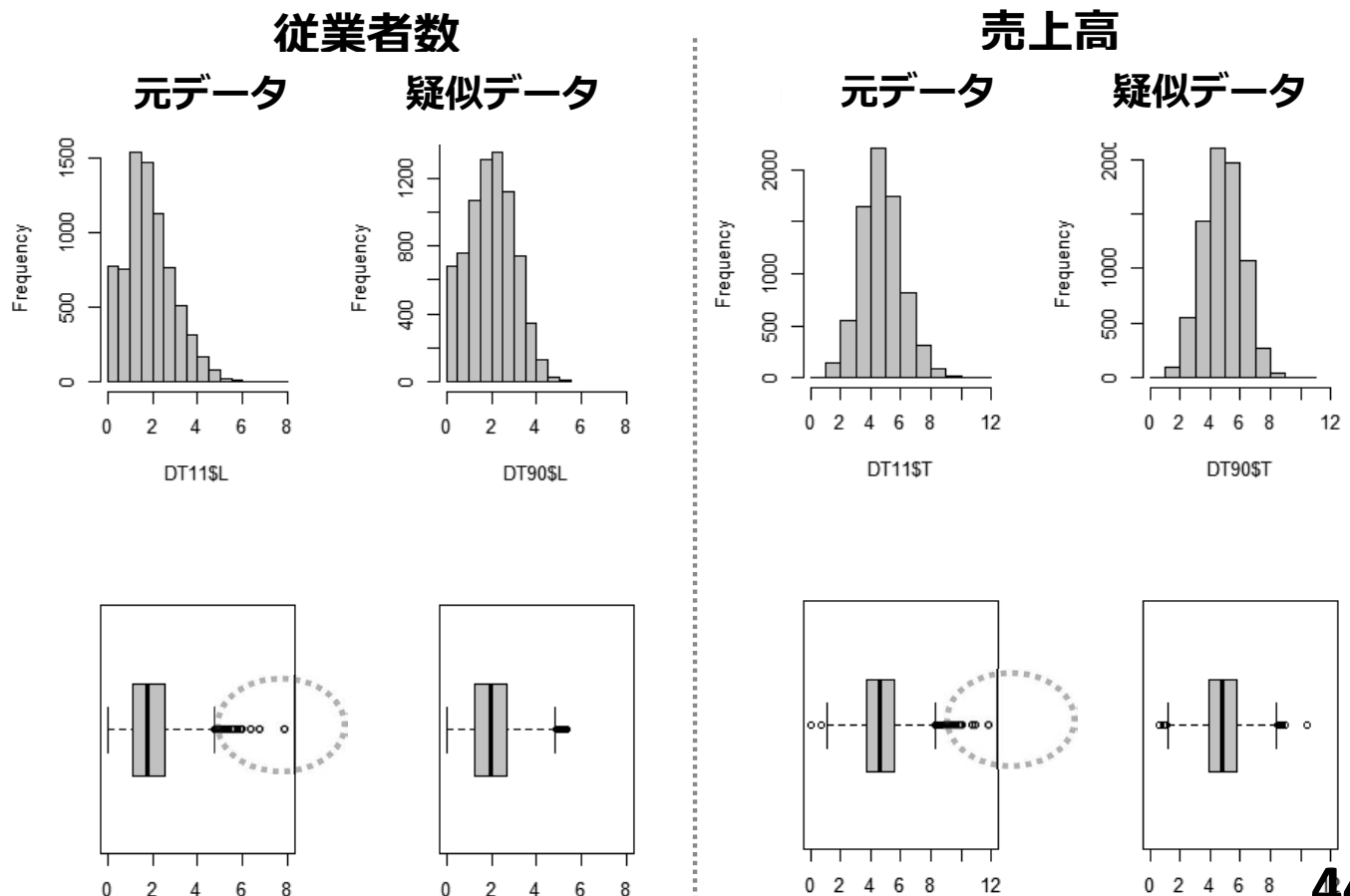
	建設業	製造業	卸・小売業	左記以外
地域 1	1065	748	645	512
地域 2	935	508	831	827
地域 3	389	357	466	275

- 上記の周辺分布に合うような（地域・産業に属する）7,558レコードを生成
- 元データから事前に推定したモデルで変数を逐次的に推測
  - (1) 地域・産業から経営組織を推測【2項ロジット】
  - (2) 地域・産業・経営組織から開設年を推測【多項ロジット】
  - (3) 地域・産業・経営組織、開設年から従業者数を推測【重回帰】
  - (4) 地域・産業・経営組織、開設年、従業者数から売上高を推測【重回帰】

※(1)・(4)：重回帰の残差から中央値(MD)、中央絶対偏差(MAD)を算出  
⇒N(MD, MAD)に従う正規乱数を付与（0で切断）

43

## 試行データの作成結果



44



# 合成データの試作②（世帯データ）

## ➤ 社会生活基本調査・匿名データ

- 時点：平成18年(2006年)
- サイズ：55,308レコード（※「世帯主」に限定）
- 変数：（※全て離散変数、適宜・カテゴリを統合）
  - ・地域（3大都市圏 or それ以外）
  - ・世帯主の性別（男女）
  - ・世帯主の年齢（10～20代、30～40代、50～60代、70代～）
  - ・教育（大卒（院卒）、高卒、その他、不詳）
  - ・ふだんの仕事（有業、有業以外、不詳）
  - ・住居の種類（持ち家、持ち家以外、不詳）
  - ・世帯年間収入（～299、300～499、500～999、1000万円～）
  - ・世帯家族類型（夫婦のみ、夫婦と子供、単身、その他）
  - ・スポーツ（1年間に行ったか）

（※欠測値（不詳）がある場合、ひとつのカテゴリとして扱う） 45

## データ作成の流れ

### ● 周辺分布の集計

（地域(2)×男女別(2)×年齢(4)）

### ● 周辺分布に合う

**55,308レコード  
を生成**

	年齢				合計
	10～20代	30～40代	50～60代	70代～	
3大都市圏					
男性	662	4388	6352	2395	13797
女性	219	644	1214	964	3041
3大都市圏以外					
男性	1422	8289	14130	6847	30688
女性	567	1495	2718	3002	7782
合計	2870	14816	24414	13208	55308

### ● 元データから事前に推定した多項ロジットモデルにより 変数を逐次的に推測

➤ 地域・性別・年齢から教育を推測

➤ 地域・産業・年齢・教育から普段の仕事を推測

➤ 地域・産業・年齢・教育・普段の仕事から住居の種類を推測

⋮

➤ 地域・産業・年齢・教育・普段の仕事・住居の種類・

世帯年間収入・世帯家族類型からスポーツを1年間に行ったかを推測

46

# 試行データの作成結果

## ➤ 教育

【レコード数】	大卒（院卒）	高卒	その他	不詳	合計
元データ	10876	24823	19078	531	55308
疑似データ	10765	24992	19044	507	55308

【構成比】	大卒（院卒）	高卒	その他	不詳	合計
元データ	0.20	0.45	0.34	0.01	1.00
疑似データ	0.19	0.45	0.34	0.01	1.00

## ➤ ふだんの仕事

【レコード数】	有業	有業以外	不詳	合計
元データ	38840	16311	157	55308
疑似データ	38767	16410	131	55308

【構成比】	有業	有業以外	不詳	合計
元データ	0.70	0.29	0.003	1.00
疑似データ	0.70	0.30	0.002	1.00

## ➤ スポーツ

（1年間に行ったか）

【レコード数】	～299万円	不詳	合計
元データ	20433	34875	55308
疑似データ	20469	34839	55308

【構成比】	～299万円	不詳	合計
元データ	0.37	0.63	1.00
疑似データ	0.37	0.63	1.00

47

# モデルの推定結果

## ➤ スポーツを1年間に行ったかを推測するモデル

### 【元データ】

Coefficients:		
	Values	Std. Err.
(Intercept)	2.07728933	0.06292023
X11	0.09264764	0.02082369
X22	-0.18700963	0.02833179
X32	-0.75566125	0.05351497
X33	-1.24322051	0.05300811
X34	-1.88517811	0.05634080
X42	-0.41095684	0.02873724
X44	-0.66800882	0.03049525
X499	-0.88185406	0.09392697
X52	0.28719127	0.02494760
X599	-0.04097486	0.16881378
X62	-0.21388003	0.02510285
X699	-0.05206236	0.12587067
X72	0.35504277	0.02522822
X73	0.67095664	0.02724565
X74	0.87789990	0.04285204
X799	0.07198791	0.06636166
X82	-0.21926362	0.02875516
X83	-0.23100521	0.03309084
X84	-0.48180454	0.02857613
X899	-0.55492871	0.06872571

Residual Deviance: 67179.48  
AIC: 67221.48

> |

### 【疑似データ】

Coefficients:		
	Values	Std. Err.
(Intercept)	2.174286813	0.06366670
X11	0.088316082	0.02083291
X22	-0.107186088	0.02836251
X32	-0.804637729	0.05382125
X33	-1.285832663	0.05337544
X34	-1.919822117	0.05667611
X42	-0.472095685	0.02903830
X44	-0.724990454	0.03086111
X499	-0.835095420	0.09577980
X52	0.259529646	0.02505617
X599	-0.297894857	0.18655347
X62	-0.223892275	0.02516053
X699	0.019661100	0.12654737
X72	0.352969143	0.02522782
X73	0.646936571	0.02723639
X74	0.950397699	0.04357236
X799	0.005754189	0.06676243
X82	-0.213276216	0.02857699
X83	-0.259202946	0.03288136
X84	-0.511732590	0.02853760
X899	-0.562102484	0.06754384

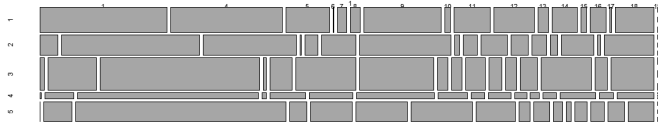
Residual Deviance: 67088.55  
AIC: 67130.55

※「住居の種類」の符号が逆になっている

48

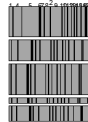
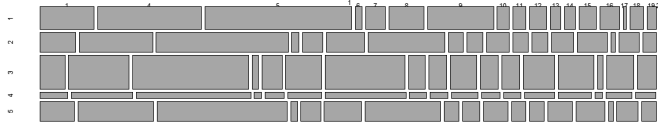
# 試行データの作成結果

table(DT06\$Z2, DT06\$Z7, DT06\$Z8)

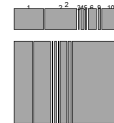
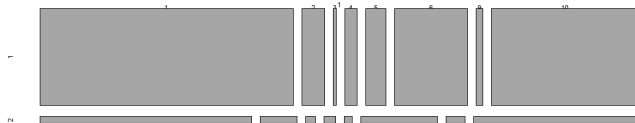


性別×企業規模×産業

table(DT07\$Z2, DT07\$Z7, DT07\$Z8)

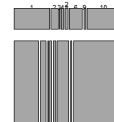
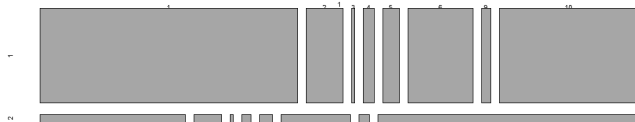


table(DT06\$Z2, DT06\$Z5, DT06\$Z6)



性別×配偶者の有無  
×就業状態

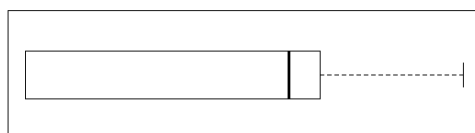
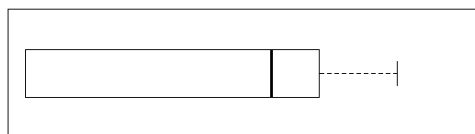
table(DT07\$Z2, DT07\$Z5, DT07\$Z6)



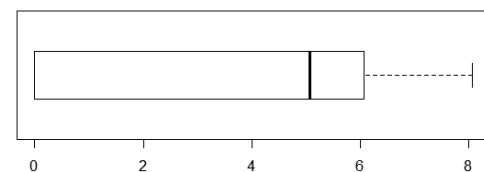
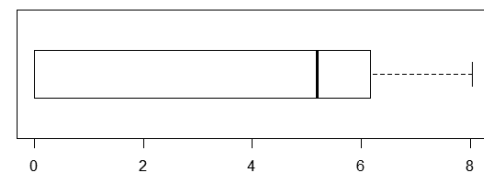
49

# 試行データの作成結果

消費支出【対数変換】



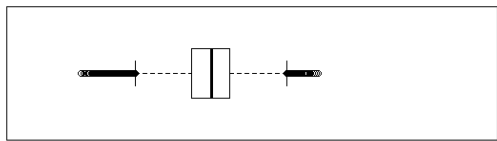
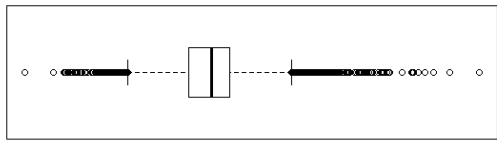
勤め先からの年間収入【対数変換】



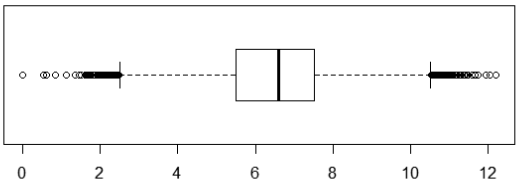
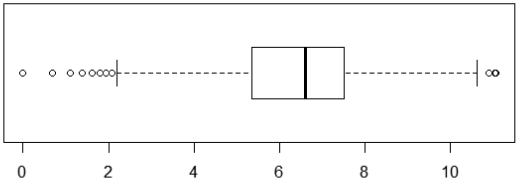
50

# 試行データの作成結果

## 住居の延べ床面積【対数変換】



## 貯蓄現在高【対数変換】



51

# 試行データの作成結果（重回帰）

## 元データ

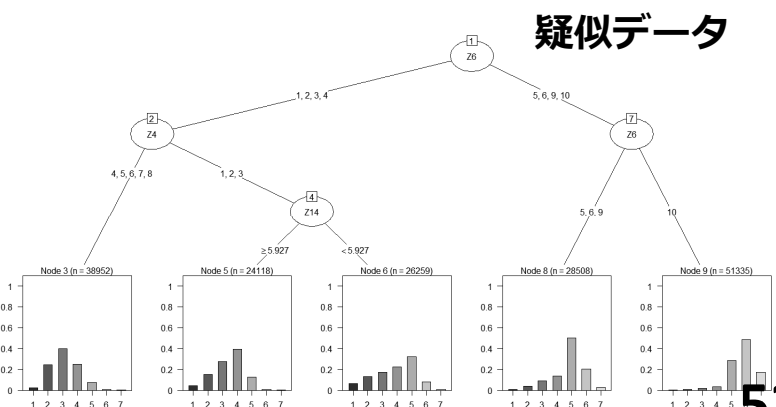
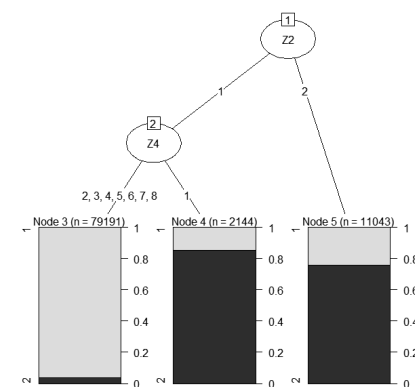
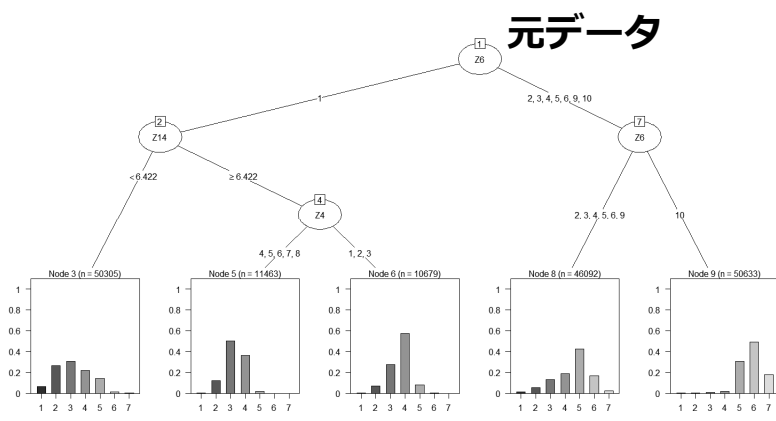
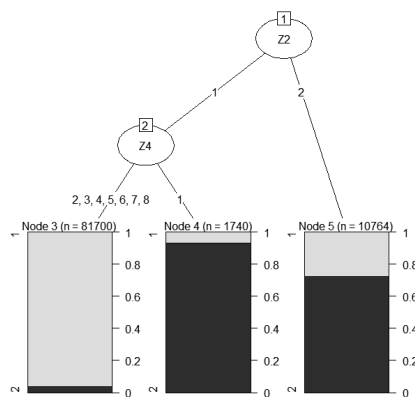
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.3042608	0.2785009	19.046	< 2e-16 ***
Z12	-0.2192660	0.1082281	-2.026	0.042771 *
Z13	0.3955380	0.0972899	4.066	4.80e-05 ***
Z14	-0.4114895	0.1140827	-3.607	0.000310 ***
Z15	0.3152534	0.1043993	3.020	0.002531 **
Z16	0.0922148	0.1019377	0.905	0.365670
Z17	0.0435758	0.1107599	0.393	0.694006
Z18	0.0803025	0.1179236	0.681	0.495892
Z19	-0.2566035	0.1046895	-2.451	0.014244 *
Z110	-0.4714310	0.1732584	-2.721	0.006510 **
Z22	0.2439973	0.0735702	3.317	0.000912 ***
Z32	0.2971309	0.1078599	2.755	0.005874 **
Z33	0.2794460	0.1058313	2.640	0.008280 **
Z34	0.1978005	0.1074093	1.842	0.065543 .
Z35	-0.1904031	0.1117055	-1.705	0.088289 .
Z36	-0.7043135	0.1274303	-5.527	3.27e-08 ***
Z37	-0.9169362	0.2066963	-4.436	9.17e-06 ***
Z42	0.1968463	0.0998924	1.971	0.048775 *
Z43	0.2713522	0.1037179	2.616	0.008892 **
Z44	0.2527787	0.1077130	2.347	0.018939 *
Z45	0.2385613	0.1180656	2.021	0.043326 *
Z46	0.2562447	0.1459841	1.755	0.079213 .
Z47	0.2971693	0.2111790	1.407	0.159374
Z48	0.4648622	0.3509302	1.325	0.185288
Z52	-0.8472326	0.0731720	-11.579	< 2e-16 ***
Z62	-0.3227735	0.0656426	-4.917	8.80e-06 ***
Z63	-0.2235163	0.1673797	-1.335	0.181754
Z64	-0.0400826	0.1044182	-0.384	0.701079
Z65	0.1848263	0.0820440	2.253	0.024276 *
Z66	-0.0404274	0.0960634	-0.421	0.673872
Z72	0.2706005	0.0881643	3.069	0.002146 **
Z73	0.4845627	0.0892113	5.432	5.60e-08 ***
Z74	0.6991325	0.1097424	6.371	1.89e-10 ***
Z75	0.9213098	0.0947511	9.723	< 2e-16 ***
Z84	0.5009367	0.0936595	5.348	8.89e-08 ***
Z85	0.5662146	0.0925616	6.117	9.56e-10 ***
Z86	0.7269883	0.2082851	3.490	0.000483 ***
Z87	1.1026582	0.1238566	8.903	< 2e-16 ***
Z88	0.4007701	0.1068353	3.751	0.000176 ***

## 疑似データ

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.694740	0.316520	17.992	< 2e-16 ***
Z12	-0.266156	0.109861	-2.423	0.015410 *
Z13	0.389798	0.098180	3.970	7.18e-05 ***
Z14	-0.493881	0.117038	-4.220	2.45e-05 ***
Z15	0.209577	0.106107	1.975	0.048254 *
Z16	0.157249	0.103282	1.523	0.127881
Z17	-0.076510	0.112507	-0.680	0.496479
Z18	-0.049666	0.119950	-0.414	0.678838
Z19	-0.385699	0.106062	-3.637	0.000276 ***
Z110	-0.568671	0.172262	-3.301	0.000963 ***
Z22	0.362801	0.073999	4.903	9.47e-07 ***
Z32	0.346913	0.112770	3.076	0.002097 **
Z33	0.248045	0.110093	2.253	0.024258 *
Z34	0.180184	0.111112	1.622	0.104884
Z35	-0.175374	0.114898	-1.526	0.126929
Z36	-0.808824	0.129556	-6.243	4.31e-10 ***
Z37	-1.286637	0.208413	-6.174	6.71e-10 ***
Z42	0.536457	0.102510	5.233	1.67e-07 ***
Z43	0.614029	0.108448	5.662	1.50e-08 ***
Z44	0.595543	0.112416	5.298	1.18e-07 ***
Z45	0.529428	0.124176	4.264	2.01e-05 ***
Z46	0.674305	0.154991	4.351	1.36e-05 ***
Z47	0.510055	0.213328	2.391	0.016807 *
Z48	0.722164	0.374500	1.928	0.053816 .
Z52	-0.754937	0.076522	-9.866	< 2e-16 ***
Z62	-0.246016	0.062775	-3.919	8.90e-05 ***
Z63	-0.130401	0.191808	-0.680	0.496601
Z64	-0.194687	0.113083	-1.722	0.085141 .
Z65	-0.027568	0.088313	-0.312	0.754923
Z66	-0.165956	0.102336	-1.622	0.104877
Z72	0.352843	0.095653	3.689	0.000225 ***
Z73	0.621744	0.096205	6.463	1.03e-10 ***
Z74	0.801333	0.116996	6.849	7.47e-12 ***
Z75	1.068675	0.100855	10.596	< 2e-16 ***
Z84	0.337668	0.087979	3.838	0.000124 ***
Z85	0.449586	0.081817	5.495	3.92e-08 ***
Z86	0.534854	0.177246	3.018	0.002549 ***
Z87	0.942788	0.119865	7.865	3.72e-15 ***
Z88	0.270012	0.101107	2.671	0.007574 **

52

# 試行データの作成結果（決定木）



53

1. 自己紹介・本日の講演概要
2. 公的統計ミクロデータの概要
3. 合成データ（Synthetic Data）の概要
4. 合成データの作成の事例

## 👉 5. 合成データの活用可能性

54

# 公的統計分野における合成データの活用可能性

- 公的統計分野への活用可能性
  - 教育・研修用データ
    - ・ データのイメージ把握
    - ・ 公的統計の学習用データ
  - 開発・検証用データ
    - ・ アプリ開発のテスト用データ
    - ・ 集計プログラムのテスト用データ
    - ・ 欠測値補完の手法検証用のデータ

## ＜資料編＞

### 資料 2-2 主催講演

「産業連関表から供給・使用表へ」

(法政大学経済学部・日本統計研究所 教授 菅 幹雄)





# 産業連関表から供給・使用表へ

法政大学経済学部・日本統計研究所  
菅 幹雄

## 産業連関表と供給・使用表

- 我が国では半世紀以上、産業連関表を推計してきました。しかし、産業連関表から供給・使用表へと移行しつつあります。この2つの表は一体、何が違うのでしょうか。
- 産業連関表は経済波及効果の計算を目的に作成されてきました。我が国では、計算結果が新聞記事になるほど、「経済波及効果」という言葉がなじみのあるものになっています。それらの計算結果はもれなく産業連関表を用いた分析（産業連関分析）によって行われたものです。
- 大阪・関西万博の経済波及効果-最新データを踏まえた試算と拡張万博の経済効果  
• [https://www.city.osaka.lg.jp/banpakusuishin/cmsfiles/contents/0000564/564186/shiryoushou\\_0129.pdf](https://www.city.osaka.lg.jp/banpakusuishin/cmsfiles/contents/0000564/564186/shiryoushou_0129.pdf)
- 我が国の中央省庁の統計関係者は多大なる熱意と労力をかけて産業連関表を推計してきました。したがって、その精度については各方面から高い信頼が寄せられてきました。国民経済計算、すなわち国内総生産（GDP）の推計にも、産業連関表が使われるようになりました。GDPは国で一番重要な経済指標です。その推計が産業連関表に大きく依存するようになったのです。
- ただし、悩みも出て来ました。産業連関表は経済波及効果の計算を目的に作成されたので、国民経済計算の推計には不都合な部分もあったのです。そこで、より国民経済計算の推計に適した表が欲しいという要望が出てきました。こうして出来たのが供給・使用表です。

# 簡単な架空のパンの例

農家 → (小麦) → 粉屋 → (小麦粉) → パン屋 → (パン) → 消費者

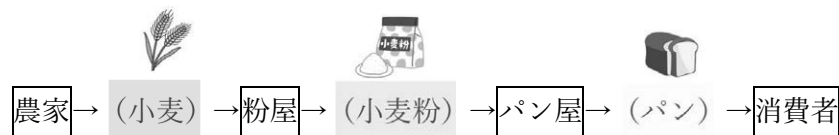


表 1 取引を示す表




	農家	粉屋	パン屋	消費者
小麦				
小麦粉				
パン				

表 2 架空の数字を入れた表

	農家	粉屋	パン屋	消費者
小麦		20		
小麦粉			50	
パン				100

表2 架空の数字を入れた表

	農家	粉屋	パン屋	消費者
小麦		20		
小麦粉			50	
パン				100

- 表2の小麦を行方向に合計します。
- そうすると小麦の行合計は農家の売上と一致します。
- 同様に小麦粉の行合計は粉屋の売上に、パンの行合計はパン屋の売上に一致します。
- これが「使用表」です。
- 行が生産物、列が産業なので、「生産物×産業表」とも呼ばれます。

表3 使用表

産業

生産物

	農家	粉屋	パン屋	消費者	行合計
小麦		20			20
小麦粉			50		50
パン				100	100
付加価値	20	30	50		
売上	20	50	100		

表3 使用表

産業

生産物

	農家	粉屋	パン屋	消費者	行合計
小麦		20			20
小麦粉			50		50
パン				100	100
付加価値	20	30	50		
売上	20	50	100		

表4 産業連関表

生産物

生産物

	小麦	小麦粉	パン	消費者	売上
小麦		20			20
小麦粉			50		50
パン				100	100
付加価値	20	30	50		
売上	20	50	100		

- 表3にある農家を小麦、粉屋を小麦粉、パン屋をパンにそれぞれ読み替えてみましょう。
- また行合計は売上と一致するので、「売上」と読み替えましょう。
- そうすると表4のようになります。これが「産業連関表」です。
- 行も列も生産物ですので、「生産物×生産物表」とも呼びます。

## 産業

表 5 供給表

	農家	粉屋	パン屋	行合計
小麦	 20			20
小麦粉		 50		50
パン			 100	100
列合計	20	50	100	170

- 一方、農家は小麦を20円、粉屋は小麦粉を50円、パン屋はパンを100円、市場に供給しています（表5）。
- これが「供給・使用表」のうちの「供給表」です。
- 左から右へみると生産物（小麦、小麦粉、パン）がどの産業（農家、粉屋、パン屋）により供給されたか、上から下へみると産業（農家、粉屋、パン屋）がどの商品（小麦、小麦粉、パン）を供給したかがわかる表となっています。

表 5 供給表

	農家	粉屋	パン屋	行合計
小麦	20			20
小麦粉		50		50
パン			100	100
列合計	20	50	100	170

表 6 副業がある供給表

	農家	粉屋	パン屋	行合計
小麦	20			20
小麦粉		50		50
パン	10		90	100
列合計	30	50	90	170

- 実際の経済はもっと複雑です。各産業では「副業」を行っているからです。
- 農家がパンを作って売っていることもあります。
- いま市場に供給されているパン100円のうち、パン屋は90円しか作っておらず、実は農家が副業としてパンを作っており、市場に10円で供給しているしましょう。
- そうすると供給表は表6のようになります。

表6 副業がある供給表

	農家	粉屋	パン屋	行合計
小麦	20			20
小麦粉		50		50
パン	10		90	100
列合計	30	50	90	170

表7 副業がある使用表

	農家	粉屋	パン屋	消費者	行合計
小麦		20			20
小麦粉	5		45		50
パン				100	100
付加価値	25	30	45		
売上	30	50	90		

- ・ 農家は粉屋から小麦粉を買わないとパンを作れません。
- ・ 農家のパンの売上10円のうち5円分を粉屋からの小麦粉購入代金、一方でパン屋の売上90円のうち45円分が粉屋からの小麦粉購入代金だとすると、使用表は表7のように変わります。
- ・ このとき表6の行合計と表7の行合計、表6の列合計と表7の売上は一致します。
- ・ これを「バランスがとれた状態」と呼びます。

## 生産物

表4 産業連関表

	小麦	小麦粉	パン	消費者	売上
小麦		20			20
小麦粉			50		50
パン				100	100
付加価値	20	30	50		
売上	20	50	100		

生産物

- ・ ところが、農家が副業でパンを作って売っても、産業連関表は表4のまま変わりません。
- ・ 産業連関表の一番上のタイトルの項目の「パン」に相当するのが、「パン屋」に加えパンを作っている「農家」が入るからです。
- ・ すなわち「生産物×生産物」の表であり、パンの供給は100円のみだからです。

表 8 2020年使用表（生産者価格，単位：兆円）

	第 1 次産業	第 2 次産業	第 3 次産業	国内最終需要	輸出	総使用 (生産者価格)
農林水産品	2	7	1	4	0	15
鉱産品、製造品、建設	3	146	58	160	57	423
その他の生産物	2	60	187	408	26	683
粗付加価値部門計	6	135	420			
国内生産額	13	349	666			

(注) 1 兆円未満を四捨五入して表示しているので、合計と内訳の計は必ずしも一致しません。  
(出所) 総務省「令和 2 年（2020 年）産業連関表」に基づき著者作成。

表 9 2020年供給表（左半分，単位：兆円）

	第 1 次産業	第 2 次産業	第 3 次産業	国内生産額
農林水産品	12	0	0	12
鉱産品、製造品、建設	0	340	9	348
その他の生産物	0	9	658	667
産出	13	349	666	1028

業 1 0 2020年供給表（右半分，単位：兆円）

	輸入、輸入品商 品税・関税	総供給 (生産者価格)	運賃・商業 マージン	総供給 (購入者価格)
農林水産品	2	15	6	21
鉱産品、製造品、建設	75	423	96	519
その他の生産物	16	683	-102	581
産出	93	1121	0	1121

(注) 1 兆円未満を四捨五入して表示しているので、合計と内訳の計は必ずしも一致しません。  
(出所) 総務省「令和 2 年（2020 年）産業連関表」に基づき著者作成。

表 1 1 2020年産業連関表（単位：兆円）

	農林水産品	鉱産品、製造品、建設	その他の生産物	国内最終需要	輸出	輸入、輸入品商品税・関税	国内生産額
農林水産品	2	8	1	4	0	-2	12
鉱産品、製造品、建設	3	148	56	160	57	-75	349
その他の生産物	2	59	186	408	26	-16	665
粗付加価値	6	133	422				
国内生産額	12	349	665				

（注）1 兆円未満を四捨五入して表示しているので、合計と内訳の計は必ずしも一致しません。  
（出所）総務省「令和2年（2020年）産業連関表」に基づき著者作成。

- ・ 経済波及効果の計算には、生産物×生産物表が必要です。
- ・ 産業連関表であれば、そのまま経済波及効果の計算ができます。
- ・ 供給・使用表の場合は、それから生産物×生産物表に変換する作業が必要になります。
- ・ そこで我が国では伝統的に直接、生産物×生産物表を推計してきました。
- ・ ところが実際に収集できるデータの大半は産業別の統計調査結果です。
- ・ 産業別の統計調査結果から作成するならば、供給・使用表の方がより自然です。
- ・ そのため海外では供給・使用表を作成し、それから生産物×生産物表に変換する方法が採用されてきました。
- ・ 職人技を駆使して、直接、生産物×生産物表を推計することは、「ものづくりを責む」我が国だったからできたと言えます。
- ・ 一方で国民経済計算では、国際的には供給・使用表からGDPを推計するようになり、その手法について国際的な基準が整備されていきました。
- ・ 国民経済計算はもともとGDPの国際比較を重視しています。
- ・ 我が国だけが特別なやり方で計算することは困るのです。
- ・ 国民経済計算の関係者から供給・使用表への移行の要請があったのは当然でした。

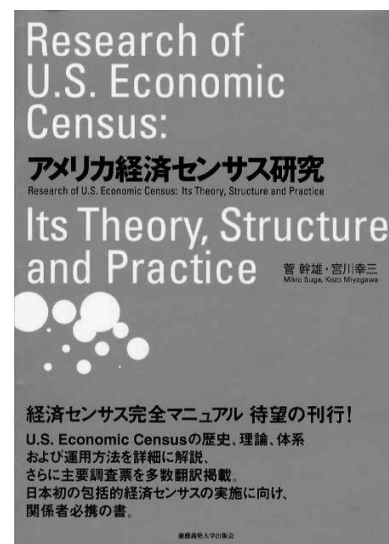
- どのような産業連関表あるいはSUTが作成可能かは、その基礎となる統計調査・資料に依存します。
- 我が国では産業連関表、欧米諸国ではまずSUTというように体系が異なっていたのは、その基礎となる産業を対象とする統計調査たる産業統計が欧米諸国とそもそも異なっていたからでした。
- 我が国において産業連関表からSUTへの移行するためには、我が国の産業統計を変えていくことが必要でした。
- 2012年には経済センサス-活動調査が開始されたことにより、これまで十分でなかったサービス産業のデータが大幅に拡充されました。
- 2019年には経済構造実態調査が開始され、それに伴い特定サービス産業実態調査、商業統計調査が廃止されました。
- 2022年には工業統計調査も経済構造実態調査に包摂されました。
- 産業連関表からSUTへの移行は、こうした産業統計の再編があって始めて実現したもののなのです。

## 経済センサス - 活動調査の開始

- 我が国では2012年に初めて調査がなされた「経済センサス-活動調査」の実現まで、全産業の経理事項を包括的に調査した全数調査（センサス）はありませんでした。
- 全産業を対象とした「事業所・企業統計調査」がありましたが、同調査は経理事項を一部しか調査していませんでした。
- それでも問題がなかったのは、我が国の経済が製造業中心であったからです。
- 当時のサービス産業は、製造業に付随するものが多く、製造業の動向と連動していました。かつそれらの産業は行政の強い規制の下に置かれていましたから、行政記録からも情報が入手できたのです。

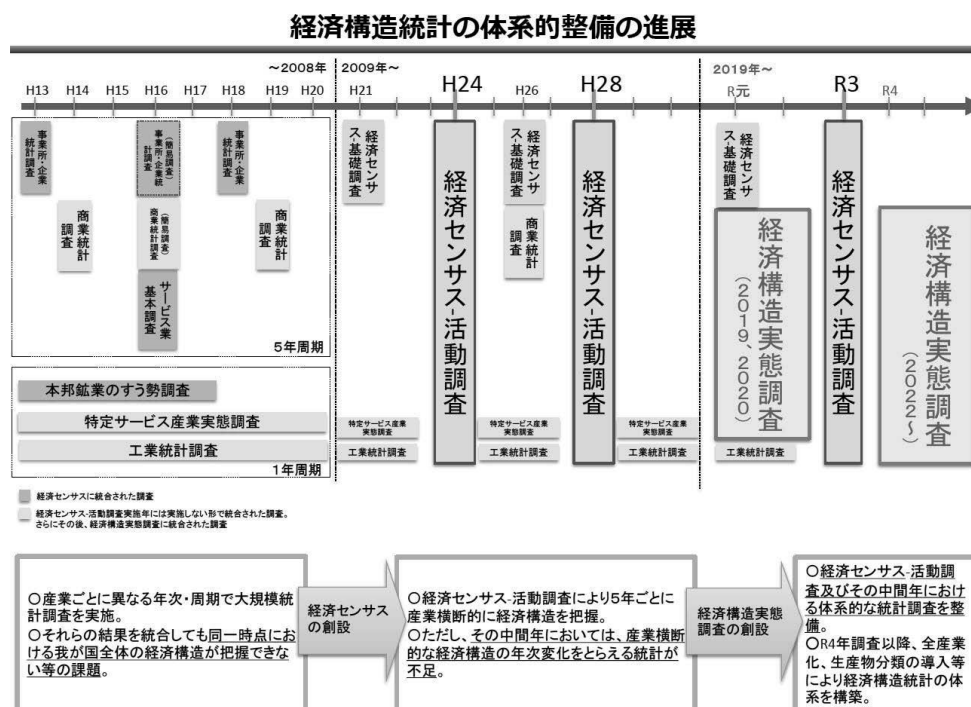


- ところが、1990年代以降、我が国の製造業が海外に生産拠点を移し（所謂「空洞化」）、かつ個人消費におけるサービスへの支出が増えると、サービス産業の統計調査の拡充が必要になりました。
- また、新しいタイプのサービス活動が増えてきました。
- そのために企画されたのが「経済センサス-活動調査」でしたが、実現までには10年近い時間がかかりました。
- 製造業では有効であった調査手法が、サービス産業には歯が立たなかったのです。製造業の場合は事業所（工場）単位で経理がなされていましたが、サービス産業の場合は事業所単位で経理がなされていない企業が多くあり、事業所単位では調査票に記入できなかったのです。
- 米国の経済センサスの調査方法に関する研究を参考に検討を重ねた末、「経済センサス-活動調査」では、事業所単位で記入可能な産業（例えば「製造業」、「商業」）については事業所単位で、ネットワーク産業など記入が難しい産業（例えば「金融業、保険業」、「運輸業、郵便業」、「通信業」）は企業単位で調査し、事後的に編集することになりました。



- ただし、これを可能にするためには企業と事業所の関係を正確に紐づける必要があり、そのために企業組織に関する調査が事前に実施され、その結果が事業所母集団データベース（ビジネスレジスター）に格納されました。
- こうして実現した「経済センサス-活動調査」はそれまでの統計調査にはない特徴があります。
- 第一に、全産業分野の売上（収入）金額や費用などの経理項目を同一時点で網羅的に把握する調査であることです。
- これまでは各種の統計資料をパッチワークのようにつなぎあわせて作成してきたので、重複や隙間（ニッチ）がある可能性が排除できませんでした。その意味で、同一時点に網羅的に把握することは、その調査結果に基づいて作成される産業連関表の精度向上につながりました。
- 第二に、事業所単位で記入可能な産業については事業所単位で、記入が難しい産業については企業単位で調査していることです。
- これにより報告者は無理のない数値を記入できるようになり、統計数値の信頼性向上につながりました。

- さらに大きかったことは、令和3年（2021年）調査では新しく策定された生産物分類に基づいて副業の生産構造を把握するための調査品目が追加されたことです。
- これにより供給表を作成するためのデータが入手できるようになりました。
- このように全産業分野の売上（収入）金額や費用などの経理項目を同一時点で網羅的に把握する調査が実現したことにより、我が国では困難とされてきたSUTの推計が可能になったのです。



出所）総務省統計局HP, [https://www.stat.go.jp/data/e-census/guide/about/sankou\\_keizai.html](https://www.stat.go.jp/data/e-census/guide/about/sankou_keizai.html)

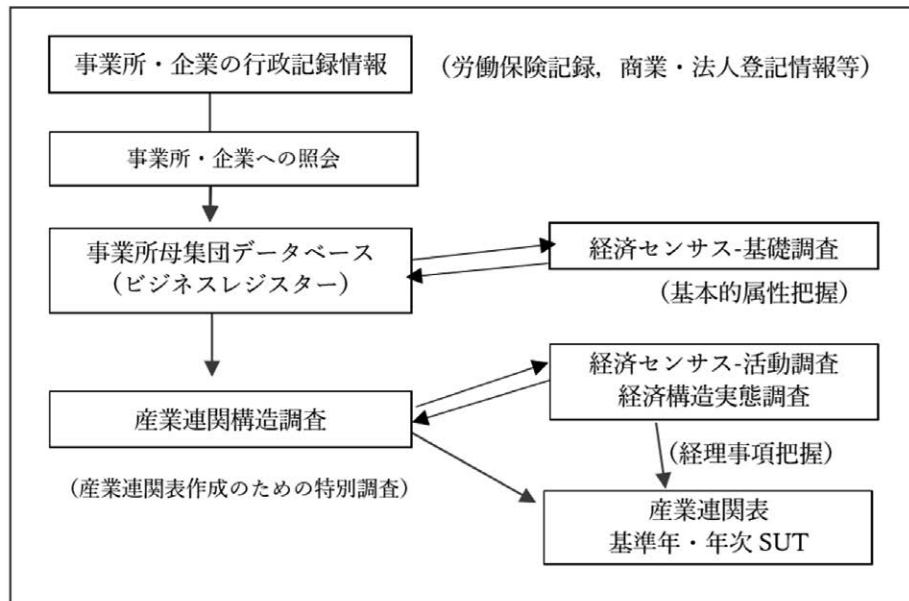


図 1 産業連関表・SUT を中心とした産業統計のしくみ

- SUTが国際標準の表であることから，欧米諸国にあわせていくような方向で産業統計の再編が進められ，2025年現在，ほぼ完了しています。
- その方向性は行政記録情報→事業所母集団データベース（ビジネスレジスター）→経済センサス-活動調査→経済構造実態調査（年次産業統計調査）という一つの流れに束ねるものであって，SUTはその終着点です。
- 2029年には2025年を対象年とする全産業のSUT，産業連関表が公表されます。
- その頃には，我が国の産業統計体系はかなりすっきりとしたものになると思います。

— 禁 無 断 転 載 —

## 「公的統計市場に関する年次レポート 2024」

～社会に生きる公的統計の共創をめざして～

2025年5月31日発行

発行所:一般社団法人 日本マーケティング・リサーチ協会  
公的統計基盤整備委員会

〒101-0044 東京都千代田区鍛冶町1-9-9 石川LKビル 2F

電話 (03) 3256-3101

FAX (03) 3256-3105

<https://www.jmra-net.or.jp>

©Copyright, 2025; JMRA, ALL Rights Reserved