

AI,IoT時代の問題解決法と人材教育

ー有効なデータ解析の背後にある基本原理ー

独立行政法人統計センター

椿 広計

第28回JUSEパッケージ活用事例シンポジウム

お品書き

- 前菜
 - 自己紹介と自組織のデータ駆動型社会における活動
 - データ解析の曙から機械学習時代まで一国内外の活動
- 主菜
 - データ解析に何を期待するのか？
 - 科学的データ解析に潜む基本原理
 - 目的変数・説明変数の選択原理
 - 説明変数の2つの階層化：分解的階層と因果的階層
 - 非線形効果と交互作用効果
- デザート
 - 統計的機械学習への途とこれからの人材教育

自己紹介と 自組織（統計センター）の データ駆動型社会における活動

職業としての統計家

データ駆動型時代を支える政府統計機関

教育用標準統計データ（SSDSE）の公開

自己紹介：自称古典統計家

- 品質マネジメント：1980-
 - 国際標準化：統計的方法の適用
 - ISO TC 69 ->207-> 176-> 69
 - **Uncertainty of Measurement**
 - Design for Environment
 - Guide 64
 - 8 Quality Management Principles
 - ISO 9000
 - Random Variate Generation
 - ISO 28640:2010
 - Process based Quality Function Deployment
 - ISO 16355シリーズ
 - TQC企業指導: 1987-1997
 - (公財) 日本適合性認定協会監理パネル:
 - マネジメントシステム第3者認証制度
 - (一社) 日本品質管理学会
 - 統計・データの質マネジメント
 - 品質管理検定の立ち上げ
- 生物統計・環境統計: 1981-
 - 医薬品許認可：中央薬事審議会
 - 臨床試験統計解析ガイドライン
 - 新薬第2調査会：循環器領域
 - 市販後調査ガイドライン：メバロチン
 - **環境計測: 1982-**
 - **国立環境研究所衛星観測チーム**
 - **オゾン層・温暖化ガス**
 - 中央環境審議会大気汚染部会PM2.5環境基準
- 公的統計:1990-
 - 統計審議会・統計委員会
 - 調査技術開発部会・農林水産部会
 - 匿名データ部会
 - (独) 統計センター
- ビジネス科学への統計応用：1997-
 - マーケティング・ファイナンスなどの実証研究支援
- 初中等数学・統計教育: 2005-
 - 教育課程部会：小中校算数・数学指導要領

Hal Varian : Google Chief Analyst

2008/10, McKinsey Quarterly

次の10年間、統計家が最も
セクシーな仕事だと
言い続けたい。

データを取る能力、データを
理解する能力、処理する能力、
データから価値を抽出する
能力、可視化し、
人に伝える能力、

それらは、次の10年間
とても
大切な技術になる。

これは、専門家レベルのこと
だけを言っているのではない。

小学生や、中高生、大学生の
教育についても言える。

なぜなら、現在私たちが、
本質的に無料なユビキタス
データを獲得したからだ。

統計家とは？：米国労働統計2017/05

米国では統計家を政府がカウント：2018年から統計家＋数学家として

- 求人状況：2017年現在全米で36,540名雇用中、今後10年で33%増加（数学家と込み）と予測、

統計家を専門職として定義

- 統計的方法を用いてデータを収集・分析し、ビジネス、エンジニアリング、科学分野など
実世界の問題のソリューション提供を支援する

統計家となるための大学・大学院教育課程が成立

- 統計か、数学の修士学位が通常必要、
最近のジョブ増加に伴い、学部レベルの職も増加。修士博士修了生は、手法開発やコンサルタントとして自立
- 統計学科修了生でも、統計家ではなく、アナリスト（金融アナリスト）や
データサイエンティストと呼ばれる職種につく人もいる。

所属組織：(独)統計センター

旧総理府統計局製表部：租庸調・報

- 1871年 太政官正院**政表課**設置（総務省統計局・（独）統計センター前身）
- 1881年 太政官統計院設置の建議と設置：初代院長 大隈重信
 - 現在ノ国勢ヲ詳明セザレバ政府則チ施政ノ便ヲ失フ
 - 過去施政ノ結果ヲ鑑照セザレバ政府其政策ノ利弊ヲ知ルニ由ナシ
- 1885年 内閣統計局設置
- **1920年 第1回国勢調査実施**
- **1946年 標本調査開始（家計調査・労働力調査）：Deming博士らが指導**
- 1947年 **統計法施行：真実性の追求**・総務庁統計局・事業所統計調査開始
- 1949年 総理府統計局**製表部**
- 1984年 **総務庁統計センター**，2001年**総務省統計センター**
- 2003年 **独立行政法人統計センター**
- 2007年 **新統計法施行：統計データは国民の資産：正しい意思決定のため**
- 2008年 **政府統計共同利用システム運用開始**
- 2009年 **第1回経済センサス実施**
- 2017年 統計改革の開始：エビデンスに基づく政策立案：オンサイト拠点ミクロデータ分析
- 2018年 **統計法・統計センター法改正,・2019年施行**

現状の把握：データの価格破壊

- 2018年6月15日閣議決定：未来投資戦略2018
- — 「**Society 5.0**」，「**データ駆動型社会**への変革」 —
 - AI,ビッグデータ，IoTの社会実装がもたらしたもの
 - **Industry 4.0**と呼ばれる産業構造の変革が世界的に進行
- ビッグデータ時代？ ⇒ **チープデータ時代**
 - 「限界費用ゼロ社会」
 - Rifkin, J. (2014) *The Zero Marginal Cost Society – The Internet of Things and the Rise of the Sharing Economy*, St Martins Pr.
 - 柴田裕之訳 (2015) 限界費用ゼロ社会<モノのインターネット>と共有型経済の台頭，NHK出版.

統計改革推進会議最終取りまとめの 全体構成（イメージ）[内閣官房2017年5月25日]

以下3スライド：内閣
官房最終とりまとめ
参考資料より抜粋

[http://www.kantei.g
o.jp/jp/singi/toukeik
aikaku/pdf/saishu_s
ankou.pdf](http://www.kantei.go.jp/jp/singi/toukeik
aikaku/pdf/saishu_s
ankou.pdf)

政策・統計の改善

1. EBPM（証拠に基づく政策立案） 推進体制の構築

- (1) 基本的な考え方
- (2) 推進の要の整備
- (3) 政策、施策、事務事業の各段階における取組

EBPMプロセスを通じた
経済統計の改善



2. GDP統計を軸にした経済統計の 改善

- (1) GDP統計の体系的整備の全体像
- (2) より正確な景気判断に資する基礎統計改善、GDP統計の加工・推計手法改善に向けた取組
- (3) 生産面を中心に見直したGDP統計への整備

経済構造の正確な把握
によるEBPMの促進

利活用促進

リソース確保

環境・基盤の整備

3. ユーザーの視点に立った統計シス テムの再構築と利活用促進

- (1) 各種データを用いた統計的分析の推進
- (2) 社会全体における統計等データの利活用の促進

4. 報告者負担の軽減と統計業務・ 統計行政体制の見直し・業務効 率化、基盤強化

- (1) 報告者負担の軽減
- (2) 統計業務の見直し・業務効率化及び各種統計の改善
- (3) 統計行政体制の見直し
- (4) 統計改革の推進の基盤強化

独立行政法人統計センターの取組み

- 国勢の基本に関する統計調査の製表
- 国又は地方公共団体の委託による統計調査の実施又は製表
- 公的統計基盤サービス
 - **統計利用者向けサービス**
 - 政府統計の総合窓口（統計ポータルサイトの運営）
 - 公的統計の二次的利用に係るサービス
(匿名データ提供、オーダーメイド集計)
 - ☆統計リテラシー向上に係る取組み → 教育用標準データセット (SSDSE)
- 統計作成・提供者向けサービス
(オンライン調査システム、事業所母集団データベース)
- 統計技術の研究

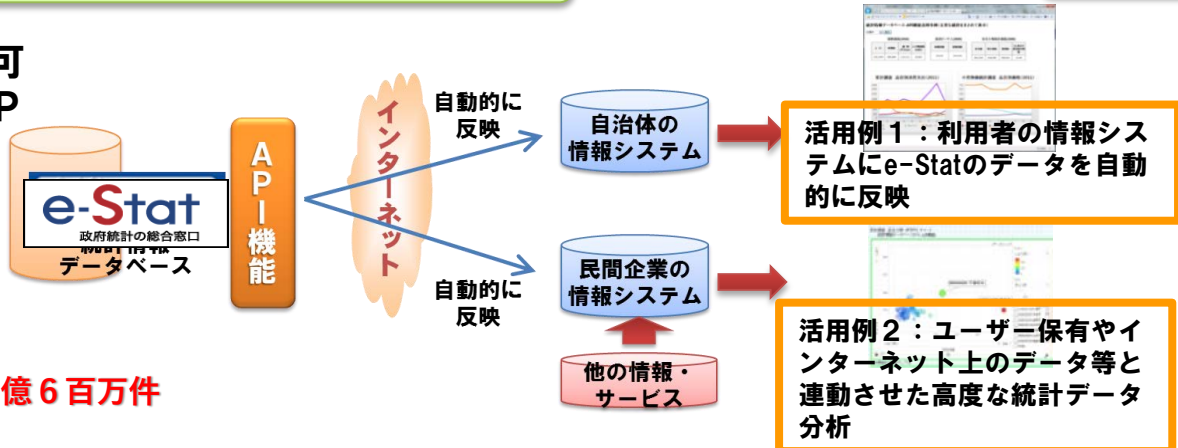
統計におけるオープンデータの高度化

統計データの提供方法を高度化し、新たな付加価値を創造するサービスや革新的な事業の創出などを支援する取り組みを、総務省統計局と連携し行っており、政府が取り組んでいるオープンデータの推進を先導。

A P I 機能による統計データの提供

2014.10.31から運用開始

統計データを機械判読可能な形式で提供するA P I 機能（Application Programming Interface）を提供中



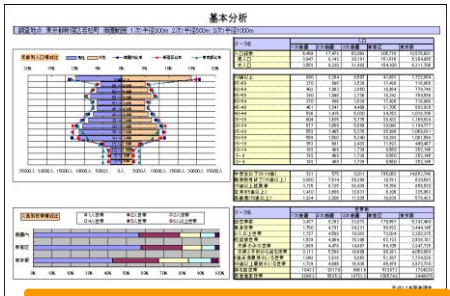
地図で見る統計（jSTAT MAP）

2015.1.20から運用開始

任意に指定したエリアによる集計や利用者が保有するデータの取り込み集計する機能などを提供

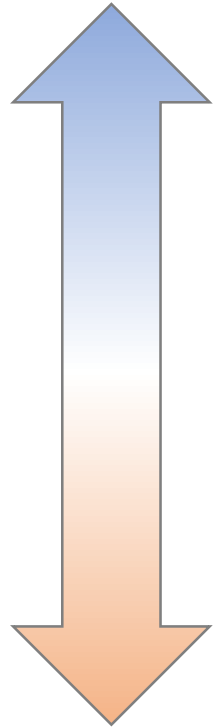


活用例1：任意に指定したエリアによる集計や、利用者が保有するデータと統計データを組み合わせ、集計結果を地図上で視覚的に把握可能



公的統計の入手と利活用

公的統計を網羅、
カスタマイズ



標準化、完備化、
ユニバーサルデザイン
(誰でも使いやすい)

- 政府統計の総合窓口 (e-Stat)

- 政府統計のポータルサイト
- 約 6 0 0 件の統計データ、約 1 3 5 万 4 千件のデータファイルを収録
- 各統計別に、地域、項目、時点等を詳細に指定してダウンロード

- 統計でみる都道府県・市区町村のすがた (社会・人口統計体系)

- 約 5 0 0 0 項目 (基礎データ + 指標) × 都道府県
- 約 8 0 0 項目 (基礎データ + 指標) × 市区町村
- 地域、項目、時点等を指定してダウンロード

- 教育用標準データセット (SSDSE)

- 1 1 1 項目 × 1 7 4 1 市区町村
- 表形式データを一括ダウンロードし、データを取捨選択して利用

教育用標準データセット (SSDSE)

Standardized Statistical Data Set for Education

- 都道府県・市区町村のすがた（社会・人口統計体系）から抽出
- 統計センターが2018/06/27 に公開 <https://www.nstac.go.jp/SSDSE/>
- 1741市区町村 × 111項目の表形式データ
 - 全市区町村（東京23区を含む）：791市, 744町, 183村, 23区 → 単純合計すると全国計に
 - 生徒誰もがどこかに所属。有名な市町村も多い。
 - 111変量抽出：完備データセットに加工（欠測値のないデータセット）
 - 複数の分野を1つのデータセットでカバー
 - 人口・世帯, 自然環境, 経済基盤, 行政基盤, 教育, 文化・スポーツ, 居住, 健康・医療, 福祉・社会保障
 - 大学データサイエンス教育, **Good Practice 共有化**への活用
- 教育用標準データの量的階層性
 - 都道府県データ・教室内データ取得（50件規模）：小中学校
 - 市町村データ, 上場企業データ（1500～2500件規模）：高校・大学
 - 定型的ビッグデータ（数万件規模）：匿名ミクロデータ：大学院
 - 画像・音声などのビッグデータ

2018/06/27 統計局・統計センターSSDSE Standardized Statistical Data Set for Education

<https://www.nstac.go.jp/SSDSE/>

• 都道府県・市町村のすがた（社会・人口統計体系）

- 全市区町村（東京23区を含む）：1741自治体
 - 791市, 744町, 183村, 23区
 - 生徒誰もがどこかに所属。有名な市町村も多い。
- 111変量抽出：完備データセットに加工
 - 人口・世帯, 自然環境, 経済基盤, 行政基盤, 教育
 - 文化・スポーツ, 居住, 健康・医療, 福祉・社会保障
- 大学データサイエンス教育, **Good Practice 共有化**への活用

• 教育用標準データの量的階層性

- 都道府県データ・教室内データ取得（50件規模）：小中学校
- 市町村データ, 上場企業データ（1500～2500件規模）：高校・大学
- 定型的ビッグデータ（数万件規模）：匿名ミクロデータ：大学院
- 画像・音声などのビッグデータ

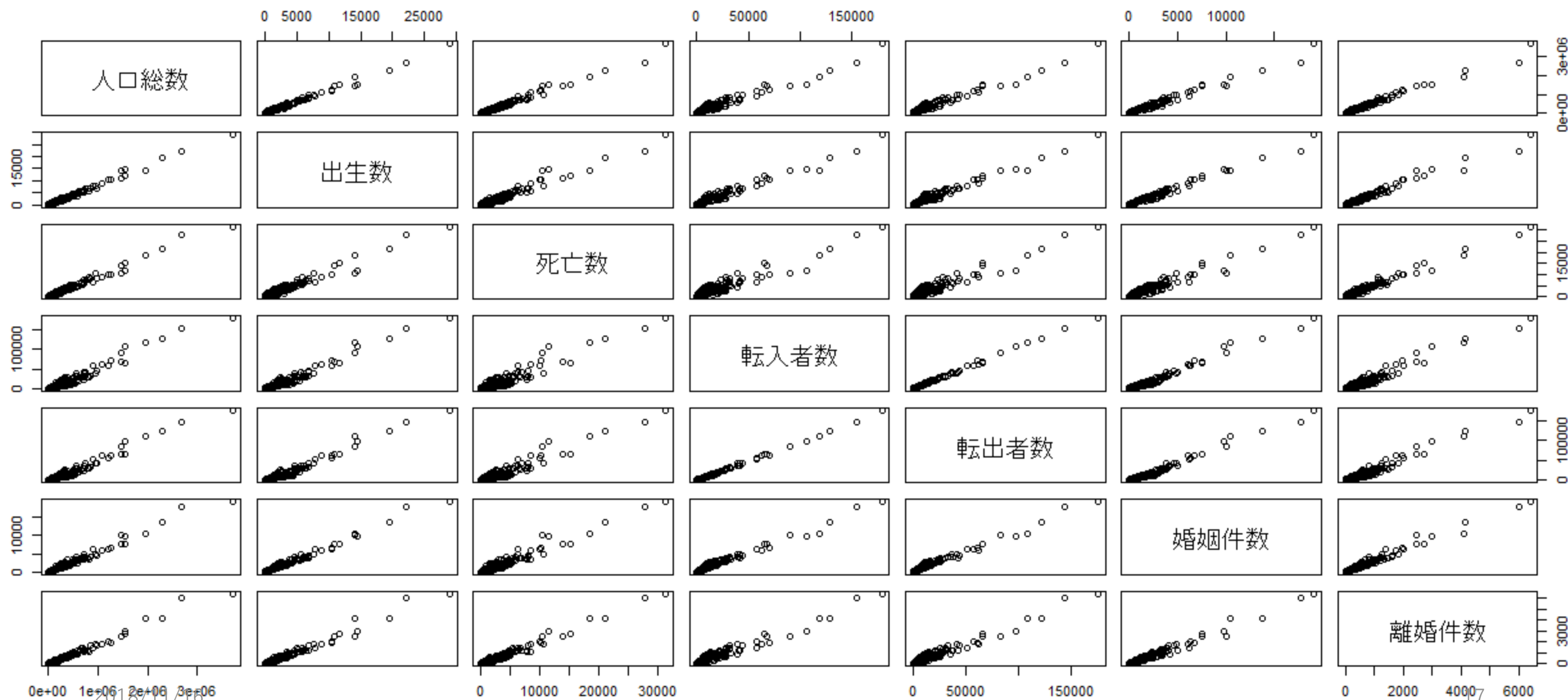
変数一覧：人口・世帯・事業所

[0]人口総数	[1]"人口総数男"	"人口総数女"	"日本人人口"	"日本人人口男"
[5]"日本人人口女"	"15歳未満人口"	"15歳未満人口男"	"15歳未満人口女"	"65歳以上人口"
[9]"15から64歳人口"	"15から64歳人口男"	"15から64歳人口女"	"75歳以上人口"	"75歳以上人口男"
[13]"65歳以上人口男"	"65歳以上人口女"	"出生数"	"死亡数"	
[17]"75歳以上人口女"	"外国人人口"	"世帯数"	"一般世帯数"	
[21]"転入者数"	"転出者数"	"単独世帯数"	"65歳以上の世帯員のいる核家族世帯数"	
[25]"一般世帯人員数"	"核家族世帯数"	"婚姻件数"	"離婚件数"	
[29]"高齢夫婦のみの世帯数"	"高齢単身世帯数65歳以上の者1人"	"事業所総数"	"事業所数農業林業"	
[33]"総面積北方地域及び竹島を除く"	"可住地面積"	"事業所数電気ガス熱供給水道業"	"事業所数情報通信業"	
[37]"事業所数建設業"	"事業所数製造業"	"事業所数金融業保険業"	"事業所数不動産業物品賃貸業"	
[41]"事業所数運輸業郵便業"	"事業所数卸売業小売業"	"事業所数生活関連サービス業娯楽業"		
[45]"事業所数学術研究専門技術サービス業"	"事業所数宿泊業飲食サービス業"			
	"事業所数教育学習支援業"			
[49]"事業所数医療福祉"	"事業所数複合サービス事業"	"事業所数サービス業他に分類されないもの"		
	"事業所数公務他に分類されるものを除く"			

変数一覧続き：事業所，従業員，財政，施設

[53]	"第 1 次産業事業所数"	"第 2 次産業事業所数"	"第 3 次産業事業所数"	"従業者総数"
[57]	"従業者数農業林業"	"従業者数建設業"	"従業者数製造業"	"従業者数電気ガス熱供給水道業"
[61]	"従業者数情報通信業"	"従業者数運輸業郵便業"	"従業者数卸売業小売業"	"従業者数金融業保険業"
[65]	"従業者数不動産業物品賃貸業"	"従業者数学術研究専門技術サービス業"		
	"従業者数宿泊業飲食サービス業"	"従業者数生活関連サービス業娯楽業"		
[69]	"従業者数教育学習支援業"	"従業者数医療福祉"	"従業者数複合サービス事業"	
	"従業者数サービス業他に分類されないもの"			
[73]	"従業者数公務他に分類されるものを除く"	"第 1 次産業従業者数"	"第 2 次産業従業者数"	"第 3 次産業従業者数"
[77]	"経常収支比率市町村財政"	"実質公債費比率市町村財政"	"歳入決算総額市町村財政"	"地方税市町村財政"
[81]	"歳出決算総額市町村財政"	"民生費市町村財政"	"土木費市町村財政"	"教育費市町村財政"
[85]	"災害復旧費市町村財政"	"幼稚園数"	"幼稚園在園者数"	"小学校数"
[89]	"小学校教員数"	"小学校児童数"	"中学校数"	"中学校教員数"
[93]	"中学校生徒数"	"高等学校数"	"高等学校生徒数"	"公民館数"
[97]	"図書館数"	"総人口非水洗化人口.水洗化人口"	"非水洗化人口"	"小売店数"
[101]	"飲食店数"	"大型小売店数"	"一般病院数"	"一般診療所数"
[105]	"歯科診療所数"	"医師数"	"歯科医師数"	"薬剤師数"
[109]	"保育所等数"	"保育所等在所児数"		

例えば人口関係の統計について、
 散布図を描くと次のようになります



SSDSEによる Good Modelling Practicesの知の全国共有

- **第1回統計データ分析コンペティション**
 - **次代を担う高校生・大学生等の統計リテラシーを向上**
 - 主催：総務省，(独)統計センター，(一社)日本統計協会
 - 後援：国立研究開発法人科学技術振興機構，
一般社団法人日本統計学会，全国統計教育研究協議会
 - 受賞論文の決定・発表：2018年10月18日（統計の日）
 - 高校生の部総務大臣賞
 - 大段利々子(広島大学附属高等学校)
 - 本当に日本の医療は危機的状況にあるのか？
 - 大学生・一般の部総務大臣賞：私の分析例よりずっと力作ぞろいです！
 - 平原幸輝（早稲田大学人間学部人間環境学科）
 - 地方創生における三つの「鍵」

「統計データ分析コンペティション」受賞者及び受賞論文の概要

～ 高校生の部 ～

受賞者	受賞論文
【総務大臣賞】 大段 利々子 (広島大学附属高等学校)	本当に日本の医療は危機的状況にあるのか？ 医療問題について、医師数、病院数と人口構成比や自治体の経済力などの関係性を分析した上で、高齢者の地域経済への貢献が重要と指摘
【優秀賞】 宮本 雨月、金山 瑠依、門脇 俊樹 (和歌山県立田辺工業高等学校)	SSDSE^注データを活用した全国学習状況調査結果との相関分析 小学生の学習への関心や取組について、家族構成や自治体の教育費等との関係性を分析した結果、大家族世帯が学習状況に良い影響を与えると指摘
【日本統計協会賞】 白石 大悟、高田 蒼大、武田 裕喜 (愛媛県立松山南高等学校)	交流人口増加による愛媛県の活性化 人口減少による経済の縮小に対して、外国人旅行者の増加を目指し、温泉旅館の利用促進を提案。様々な公的統計を活用し、経済波及効果も推計
【特別賞】 伊藤 寛子 (渋谷教育学園幕張高等学校)	機械学習による15歳未満人口の推定 機械学習により人口統計データを分析することを試み、その過程を詳細に記載

注 (独)統計センターが提供している「教育用標準データセット(Standardized Statistical Data Set for Education)」。
統計教育等での利用に供するため、市区町村別の主要データを簡易な表形式に編集したもの。

「統計データ分析コンペティション」受賞者及び受賞論文の概要

～ 大学生・一般の部 ～

受賞者	受賞論文
【総務大臣賞】 平原 幸輝 (早稲田大学人間科学部人間環境科学科)	地方創生における三つの「鍵」 地方創生のコアである人口増減について、「大学進学率」、「人口当たり医師数」、「労働力人口率」の3つの指標が強い影響力を持っていると指摘。これらの指標に基づき、各市町村を類型化
【優秀賞】 小野島 昂洋 (早稲田大学大学院教育学研究科)	人口規模によって異なる保育所数・保育所在所児数・定員充足率の関係 保育所数、利用者数、定員充足率は市町村の規模によって、その関係性に差異があることを確認した上で、政策立案においてはこの差異を考慮する必要があると指摘
【日本統計協会賞】 池田 泰成、柴辻 優樹、鶏内 朋也、 石川 貴啓、佐野 岳史 (慶應義塾大学リーディングプログラム)	地方創生に向けた東京一極集中是正のための定量的都市圏選定指標の提案 首都機能を他の都市へ移転することを想定し、地理、経済、生活に係る指標を使って、移転にふさわしい都市を具体的に選定
【特別賞】 小野 恵子、宮内 はじめ、白松 俊、 河口 信夫、五十嵐 康伸 (国際基督教大学社会科学研究所、Code for Nagoya・OSGeo財団日本支部、名古屋工業大学大学院工学研究科、名古屋大学大学院工学研究科、パーソルキャリア株式会社)	日本の全市町村における人口の自然増減の分布と説明要因 人口の自然減少について、地理的・時間的な分布を明らかにするとともに、高齢者率、若年女性比、女性・子ども比、都市圏かどうかの影響していると指摘

データ解析の曙から 機械学習時代まで

データ解析の登場

日科技連多変量解析研究会

統計的機械学習の時代へ

Tukeyとデータ解析

Future of Data Analysis & EDA

- データ解析誕生
 - Tukey, 1962
- 裁判官の立場の推測統計
 - 仮説検証型統計推測
- 刑事の立場のデータ解析
 - Tukey, 1977
- Bell 研究所の指導
 - C言語からS言語へ
 - データ解析環境

J.W. Tukey, J.H. Friedman and M.A. Fisherkeller
Stanford Linear Accelerator (1973)
26:00 minutes



<http://stat-graphics.org/movies/prim9.html>

日本の産業界でのデータ解析の曙 奥野・久米・芳賀・吉沢(1971)「多変量解析法」

- MA研
 - 日本科学技術連盟
 - 奥野忠一・芳賀敏郎・吉澤正
 - 産学若手が集うデータ解析道場
- 多変量解析事例集
 - データ解析の知の共有
 - データの共有
- 芳賀の対話的データ解析
 - 多変量解析ソフトウェア
 - 日科技研：JUSE-MA



芳賀敏郎記録集より
<http://haga.sakura.ne.jp/>

モデリングから統計的機械学習 (AI) 回帰予測問題を中心に

統計モデル接近

2段階接近：現象解釈⇒予測

- モデリング戦略
 - Shewhart:統計的管理戦略：系統変動＋偶然変動
 - 外れ値要因探索：系統変動の発見
 - 田口：系統変動の分類（制御因子と標示因子）
 - 系統変動の偶然変動化：確率分布課程の正当化
 - Fisher/・Neyman:無作為観測（割付・抽出）
 - 田口：偶然変動をSimulate（ノイズ因子）
- 正規・線形モデルの拡張
 - 系統変動モデルの拡張
 - 非線形性モデル：メカニズム情報の考慮、変数変換
 - ノンパラメトリックモデル：探索的変換
 - 連立方程式モデル：因果情報の考慮
 - 潜在変数モデル：観測されていない原因系の考慮
 - 平均値以外のモデル：バラツきのモデル、順位モデル
 - 偶然変動モデルの拡張
 - 分散関数導入：確率変数の平均と分散との関係
 - 誤差の階層性

統計的機械学習とブラックボックス接近

一段階接近：予測損失関数最適化

- 探索的データ解析
 - ロバスト統計
- データマイニング：データからの知識発見
 - 非ブラックボックス：第2世代人工知能
 - CART,推論エンジン⇔樹形モデル
- 第3世代人工知能
 - 統計的機械学習の数理統計的戦略
 - 強化学習・敵対的学習⇔逐次実験計画
 - ブースティング：推定量の統合
 - Random Forest
 - 正則化⇔ベイズモデルと縮小推定
 - カーネルトリック⇔計量的多次元尺度構成
 - サポートベクターマシン
 - 深層学習⇔非線形多層潜在因子モデル
 - 多層ニューラルネットワーク

データ解析に 何を期待するのか？

目的の分類：発見・分析・最適化

モデリング（分析機能）のプロセス

デミング・石川モデルのSHINKAの方向性

データサイエンスによる問題解決

モニタリング機能
利用プロセス
リアルタイム可視化

Do

PDCAサイクル
顧客接点も
プロセス管理対象
→顧客対応工程が
価値の源泉
サービス科学

Plan

最適実装

問題発見機能
Check ビッグデータによる問題発見加速
平均値予測よりも外れ値発見

あるべき姿と
実際とのずれ
What, Who,
When, Where,
How

問題提起

QCストーリー
自動改善（調整）システム
システム改善は人間の役割
自律改革は人の役割
目的の追加，制約の強化

Action = 対策立案

最適数理計画機能
多目的制約付き最適化
+ 実装効果確認の予測

仮説提示

Data Consolidation
技術の活用
計るべきものを
どう自動結合するか

量的調査計画
最適実験計画
数値実験計画
(直接関係ないが
データの原価低減)

情報収集 →
情報創成

分析

モデリング機能
因果モデルの機械学習
シミュレーションによる予測

データサイエンス（実は技術）対 統計科学

Breimanの認識

- 技術：予測パフォーマンスに対する最適化ならば**外注可能性**が高い
 - 統計的方法の技術利用＝統計的機械学習（AI）の優位性は顕著
 - 予測・制御・決定の一段階最適化
 - 意図的にブラックボックス化された機械学習方式??
- 科学＝解釈：データ取得計画に基づくモデリングと現象理解の意義
 - **2段階接近：モデル化（メカニズム理解）の後の最適化（メカニズム利用）**
 - 技術モデルのデータ同化
 - 法則（構造）と誤差（測定方法＋測定環境）の記述
 - いざとなったら最適実験計画（日本のお家芸としてMITからかつて注目）
 - 少数データでビッグデータと同等の情報量確保⇒ビッグ情報量
 - **集まるデータ（価格破壊進行中）**だけでは本質的に因果分析は不可能
 - **集めるデータ＋実験的介入**でこそ因果関係を読み切らなければならない場合
 - 確実に社内エキスパートの役割

設計科学の文法に組み込まれる Cox & DonnellyのIdeal Process

- Shewhartの問題発見
 - 外れ値の発見と要因分析
 - 本質的な外れ値とデータ分析の未熟さによる外れ値
- 分析のプロセス
 - 研究すべき問題，仮説の定式化
 - 関連するデータの探索と適切なデータを採取する研究の計画と実施
 - データ解析
 - 適切な意思決定に繋がる結果の解釈
- 最適化
 - 前工程の分析結果が最適化に寄与しやすい出力か？
 - 応答曲面法（2次形式の最大化）

Explanatory Analysis：モデルビルディング (椿,筑波大多変量解析第1講義資料2001)

- 現象のメカニズム推論の意義
 - 統計科学へ：Fisherの科学的推論への回帰
 - 統計により、与えられる付加価値とは、現象に対するより定量的な理解
 - データの変動を「説明」するモデルを当てはめ、作業仮説を成長させよ！
 - モデルの要素として以下が有ることを認識せよ
 - 構造： 普遍的構造部分
 - 測定： 技術的限界が突破されれば変動しうる構造
 - 標示： 採取したデータに特有の構造
 - 標示・測定部分に含まれる母数は、モデル利用の場に応じて再チューニングせよ
- 一般化（外挿）可能性への挑戦こそ、モデル思考の意義

科学的データ解析に潜む基本原理

目的変数・説明変数の選択原理
説明変数の2つの階層化：分解的階層と因果的階層
非線形効果と交互作用効果

統計的問題（解析目的）とその分解 目的変数（結果変数）の選択

- 統計的問題
 - 入力が不適切で狙いからずれる
 - 入出力関係を使って出力を調整
 - そもそも出力データがばらつき,
 - 許容範囲に入らない
 - バラつきの原因を探る
- 統計的問題解決
 - 原因に対して対処
- 目的（問題）の分解
 - 利益が少ない
 - 利益 = 売上 - コスト
 - 売上 = コスト × (売上/コスト)
 - コスト削減問題
 - コストの割に売り上げが出ない
 - 出生数が少ない
 - 高齢化, 未婚化, 少子化
 - 出生数 = 人口 × (出生可能年齢者数/人口) × 婚姻件数/出生可能年齢者数 × 出生数/婚姻件数

SSDSEから20変数抽出 分析戦略立案：仮説的因果階層化

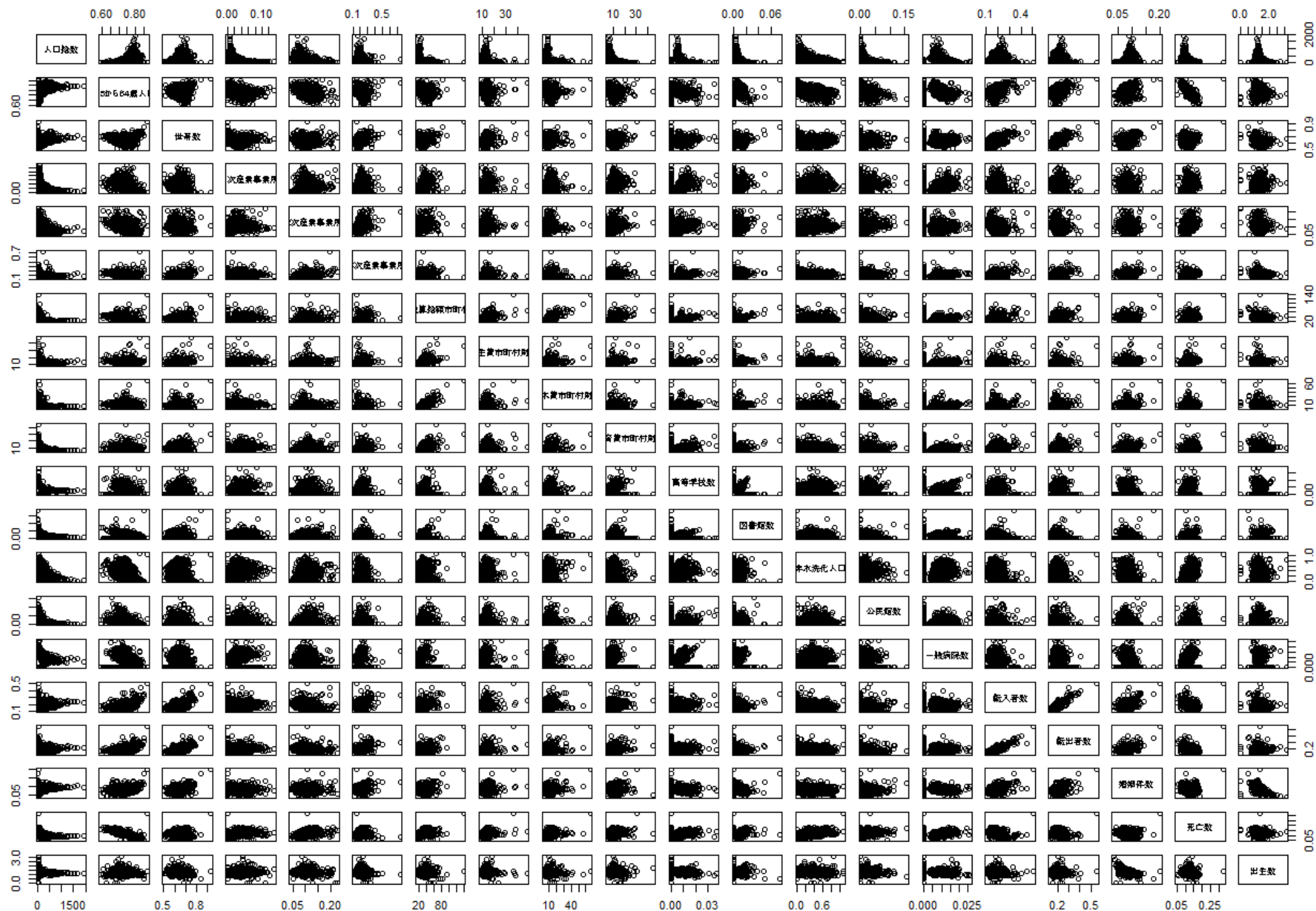
- 第0階層：人口総数
- 第1階層：自治体の現状変数（率）
 - 人口特性：15から64歳人口(労働力人口), 世帯数
 - 経済特性：第1次産業事業所数, 第2次産業事業所数, 第3次産業事業所数
- 第2階層：自治体の資金特性（率）
 - 収入：歳入決算総額市町村財政
 - 支出：民生費市町村財政, 土木費市町村財政, 教育費市町村財政
- 第3階層：自治体内の施設オペレーション
 - 高等学校数, 図書館数, 非水洗化人口, 公民館数, 一般病院数
- 第4階層：自治体の第1次パフォーマンス：転入者数, 転出者数, 死亡数
- 第5階層：自治体の第2次パフォーマンス（率）：婚姻件数
- 第6階層：第3次パフォーマンス（率）：出生数

主要な説明変数の選択：Efficiencyの分析

人口総数, 婚姻件数

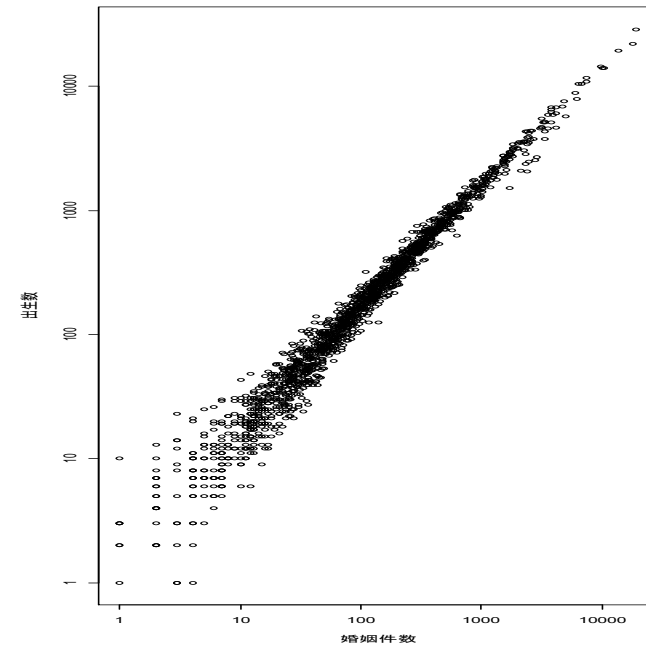
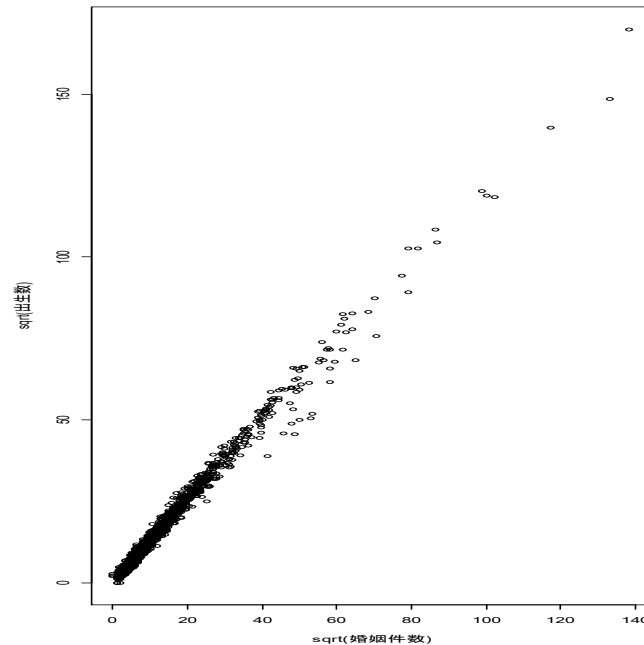
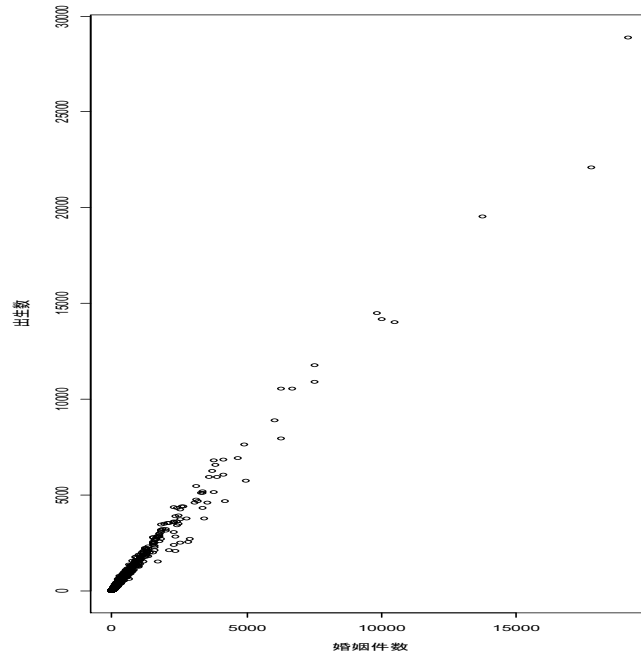
- 第0階層：人口総数
- 第一階層から第5階層変数を人口総数で割り、人口当たり変数化
 - 人口0あるいは労働力人口0の自治体は除外
 - 福島第一原発周辺自治体
 - 人口0: 富岡町、大熊町、双葉町、浪江町
 - 労働力人口0: 飯館村
 - 歳入0: 福島県葛尾村
 - 近似的に比例関係にあることを前提
 - 御用とお急ぎ向き：統計的機械学習利用の際には有力
- 出生数は、婚姻件数当たり変数加工
 - 直近婚姻件数が0の東京都御蔵島村は
 - 第5階層：自治体の第2次パフォーマンス：婚姻件数
 - 第6階層：第3次パフォーマンス：出生数
- 人口総数で基準化せず、比例関係を微修正する戦略も有力

20変数の
平方根変換
多変量連関図



回帰分析と残差分散の均一性仮定

主要な説明変数(信号因子)を知っているからこそ
婚姻件数と出生数の散布図：原変数・両平方根・両対数



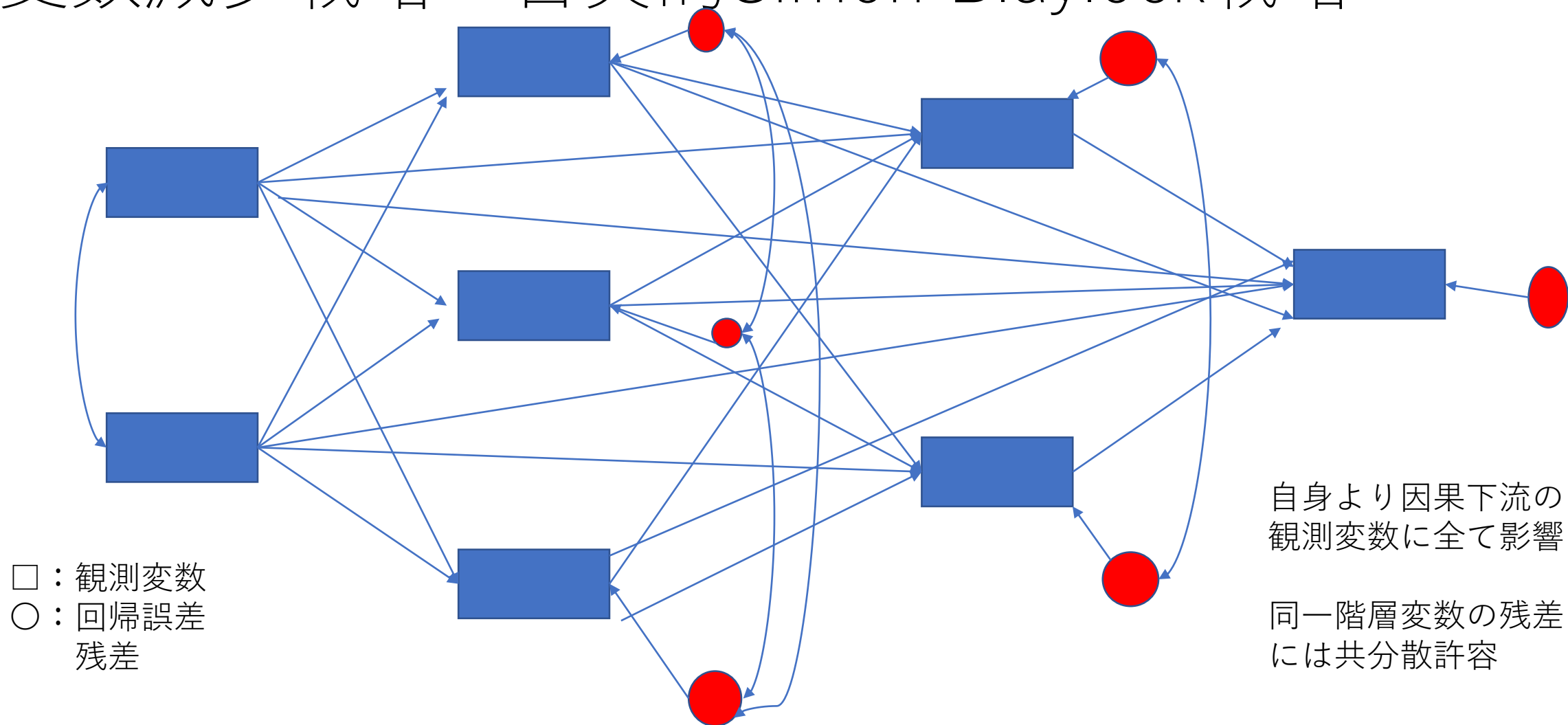
両平方根変換 $\Rightarrow$ 等分散性 $\Rightarrow$ 出生件数が擬似ポアソン変数

分散が平均に比例する

品質工学のSN比の2段階設計が上手くゆくための要件でもある

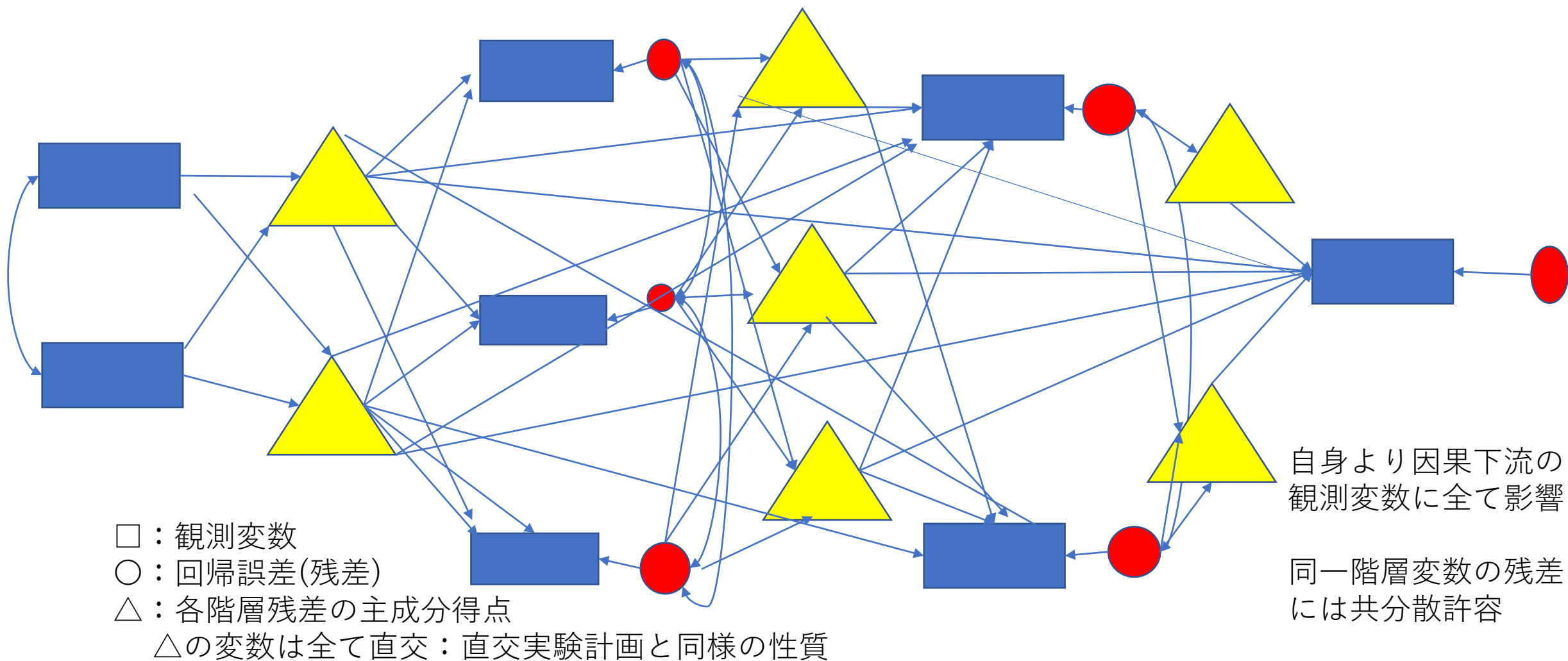
以後kan簡単のため、効率性指標も全て平方根変換

仮説的因果構造に対する飽和モデルからの 変数減少戦略：古典的Simon Blaylock戦略



修正飽和モデル戦略：変数直交化

第0層と各層偏残差の主成分回帰



第0層（人口）⇒第1層（労働人口率、事業所率）

- 人口総数と目的変数との標準偏回帰係数

目的変数	標準偏回帰係数
労働人口率	0.491
世帯率	0.177
第1次産業事業所率	-0.481
第2次産業事業所率	-0.364
第3次産業事業所率	-0.186

- 残差の主成分負荷量

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分
労働人口率	0.656	0.304	-0.367	0.583	0.056
世帯率	-0.571	0.714	-0.179	-0.137	0.336
第1次事業所	-0.612	-0.032	0.622	0.481	0.074
第2次事業所	-0.453	-0.706	-0.458	0.123	0.268
第3次事業所	-0.800	0.164	-0.390	0.137	-0.403

因果上層からの回帰残差主成分負荷量

第2層：自治体の歳出入

	第1主成分	第2主成分	第3主成分	第4主成分
歳入率	-0.956	0.039	-0.143	0.253
民生費率	-0.626	-0.691	0.355	-0.059
土木比率	-0.830	0.055	-0.522	-0.187
教育費率	-0.659	0.531	0.527	-0.076

第3層：自治体内の施設

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分
高校率	-0.699	0.024	-0.291	-0.202	0.620
図書館率	-0.196	-0.880	-0.021	0.431	0.031
非水洗人口率	-0.580	0.482	-0.109	0.613	-0.209
公民館率	0.253	-0.052	-0.946	-0.039	-0.192
一般病院率	-0.698	-0.197	0.044	-0.442	-0.527

主成分回帰

出生率の回帰分析結果
標準偏回帰係数
(相関係数)

第4層：転出・転入・死亡率

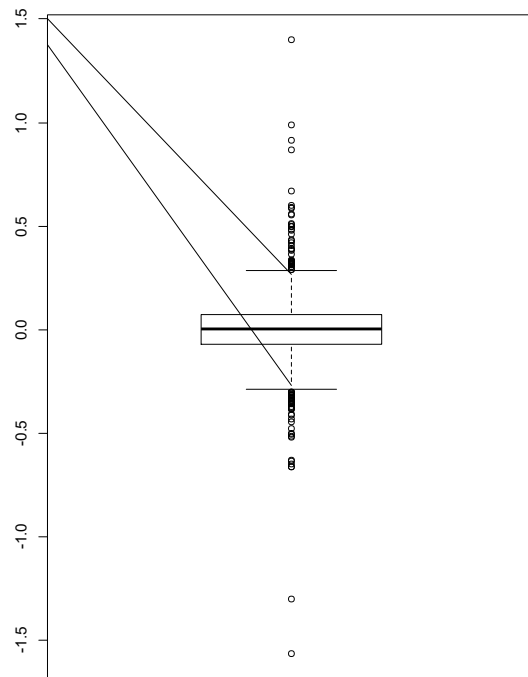
	第1 主成分	第2 主成分	第3 主成分
転入率	-0.903	0.035	0.365
転出率	-0.915	0.184	-0.359
死亡率	0.206	0.977	0.055

	婚姻率	出生率
人口総数	0.287***	-0.066***
PCA11	-0.018	0.187***
PCA12	0.304***	-0.052**
PCA13	-0.279***	0.148***
PCA14	0.183***	0.022
PCA15	0.050**	-0.124***
PCA21	-0.117***	-0.047*
PCA22	-0.044*	0.030
PCA23	-0.045*	0.042*
PCA24	-0.028	0.030
PCA31	-0.008	0.006
PCA32	-0.056**	-0.018
PCA33	-0.032	0.032
PCA34	0.028	0.011
PCA35	-0.037	-0.032
PCA41	-0.213***	-0.096***
PCA42	0.054**	-0.081***
PCA43	0.077***	0.054**
婚姻率残差	---	-0.494***
寄与率	0.368	0.351

チャンス発見は残差から

出生率
出生件数/婚姻件数の平方根
残差ボックスプロット

残差ベスト10



都道府県	市町村	人口	婚姻率	出生件数/ 婚姻件数
北海道	島牧村	1499	0.0013	10.0
福島県	檜葉町	975	0.0391	1.9
宮崎県	諸塚村	1739	0.0024	6.5
宮崎県	五ヶ瀬町	3887	0.0016	7.7
島根県	海士町	2353	0.0034	5.0
北海道	黒松内町	3082	0.0025	5.3
山梨県	鳴沢村	2921	0.0030	5.0
北海道	北竜町	1981	0.0032	4.7
奈良県	野迫川村	449	0.0047	3.0
和歌山県	北山村	446	0.0152	1.7

階層化原理：群間変動と群内変動

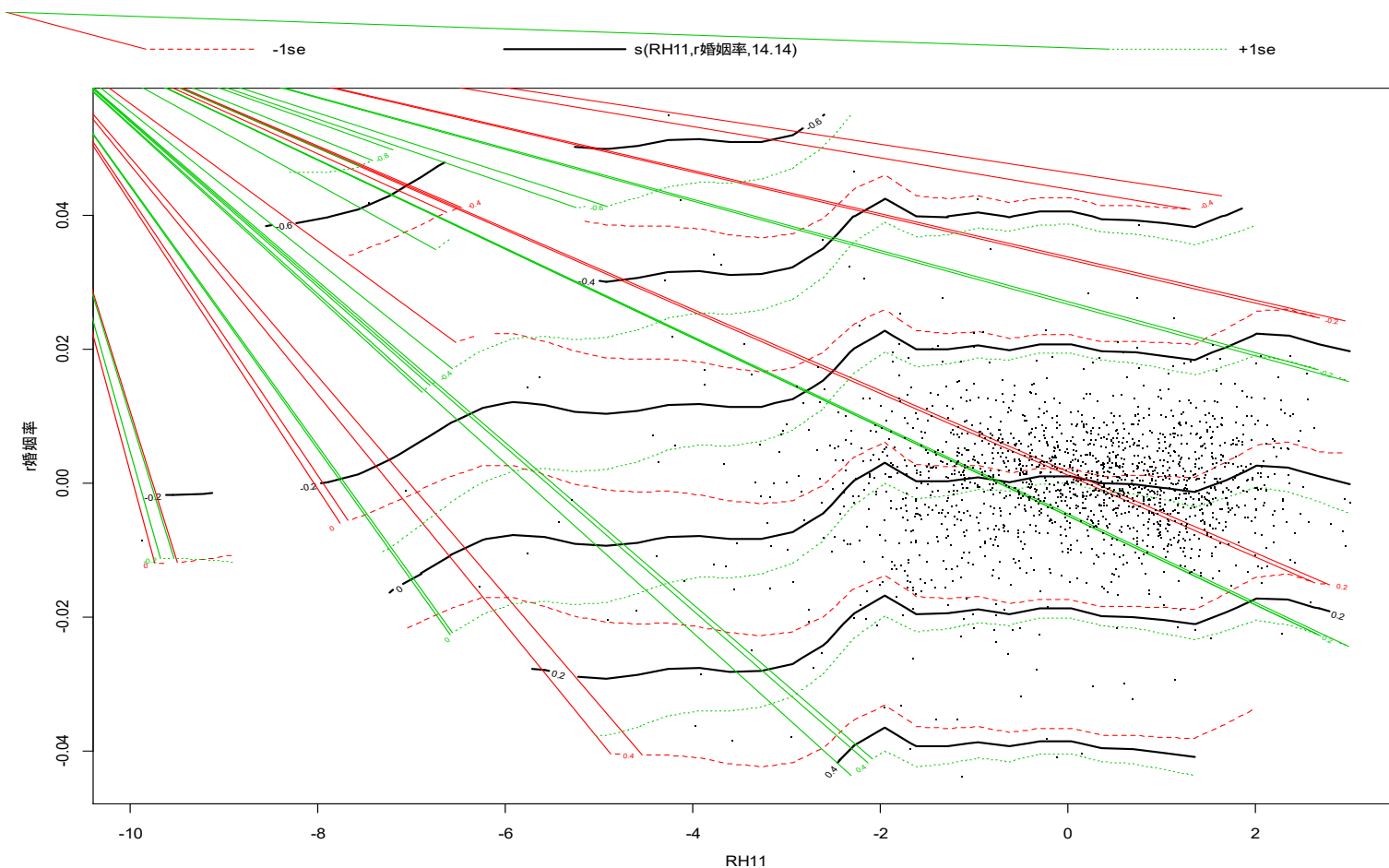
奥野・芳賀分割型回帰orマルチレベルモデ～階層ベイズモデル

- 機械学習のための入力変数？
 - 都道府県と市区町村
 - 都道府県層別
 - 都市部・山間部層別等
 - 説明変数の分解：
 - 都道府県平均値 +
(原説明変数-都道府県平均値)
 - 2変数を説明変数に投入！
 - 2変数とも有意となることが殆ど
 - 労働力人口率
 - 人口規模の効果
 - 第1次、第2次、第3次事業所率
 - どちらの変数も負で有意
 - 人口規模により効果低減

- 労働力人口率の予測

	係数	標準 誤差	T 値	P値
切片	0.710	0.002	351.5	0.00
県平均 人口	1.841 E-04	0.092 E-04	20.0	0.00
県平均 からの 残差	7.663 E-05	0.447 E-05	17.2	0.00

非線形性と交互作用への20世紀末の対処原理 一般化加法モデル (GAM, MARS)：平滑化原理



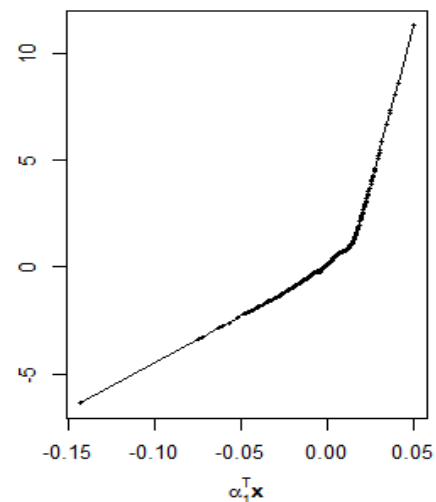
非線形項導入は統計的に有意
回帰曲面をスプライン関数で近似
RH11の効果か？
交互作用は有意でない

Projection Pursuit Regression：射影(尺度) 追跡

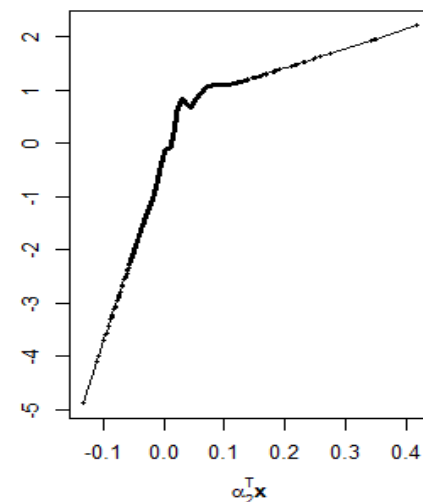
	第1尺度	第2尺度	第3尺度	第4尺度
人口総数	0.000	0.000	0.000	0.000
PCA11	0.003	-0.005	0.232	-0.027
PCA12	-0.004	0.014	-0.016	-0.076
PCA13	0.000	0.000	-0.230	-0.001
PCA14	-0.002	0.012	-0.250	-0.026
PCA15	0.000	-0.009	0.389	-0.021
PCA21	0.002	-0.009	-0.022	-0.018
PCA22	0.001	-0.023	-0.016	-0.082
PCA23	0.001	0.008	0.011	0.077
PCA24	0.002	-0.005	-0.118	-0.073
PCA31	0.000	0.005	-0.094	0.058
PCA32	0.000	-0.007	0.177	0.090
PCA33	0.001	-0.002	-0.049	0.027
PCA34	0.002	-0.020	0.244	0.013
PCA35	0.000	0.003	-0.278	-0.002
PCA41	0.000	-0.010	0.054	0.003
PCA42	-0.003	-0.004	0.371	-0.128
PCA43	-0.002	0.034	-0.498	-0.003
婚姻率残差	-1.000	-0.999	0.347	0.972
係数	0.127	0.062	0.047	0.049

変化点と非線形性の同時抽出
3層ニューラルネットを含むモデル

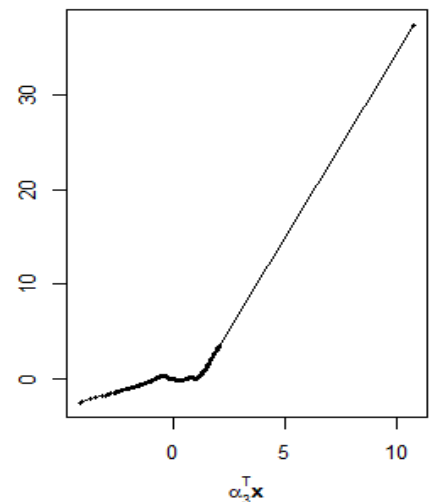
term₁: $\hat{\beta}_1 = 0.127$



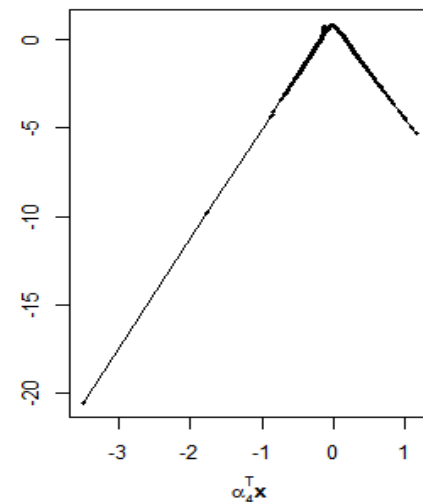
term₂: $\hat{\beta}_2 = 0.0621$



term₃: $\hat{\beta}_3 = 0.0474$



term₄: $\hat{\beta}_4 = 0.049$



統計的機械学習への途と
これからの人材教育

ビッグデータと統計的機械学習の誤解

- **AI = Machine Learning**は画期的な予測理論？：No
 - 予測に用いる統計的機械学習→回帰分析・層別分析の自動化(自動的学習)
 - **回帰予測**とは何か？→予測に用いる情報を与えたときの被予測情報の条件付平均値予測
 - 良い**層別**とは何か？→群内のばらつきを最小化
 - **最適意思決定**とは何か→**損失関数**を平均的に最小化
 - 機械学習が採用している幾つかの原理とは何か？→単純だが利用者が意識すべき
- ビッグデータから**Nowcast**や**行動の可視化**は完璧に近づく：Yes
- 既存ビッグデータからの**将来予測**や行動選択は完璧に近づく：No
 - どんなモデルを用いても将来の不確かさを超えた予測は不可能
- **行動とその結果の評価モデル（損失関数・価値関数）が存在する場合**
 - **既存ビッグデータではなく自律的実験データを創造的に大量生成・分析**することで
 - 行動選択は既存の最適行動を上回るパフォーマンスとなる；**Yes**
 - **強化学習**：人工知能が自律的にPDCAサイクルを回すことができれば
 - 革新的技術が生まれる可能性⇒数値シミュレーション＋品質工学（制御因子vs敵対的因子）
 - **敵対的学習**

機械学習の幾つかの主要原理

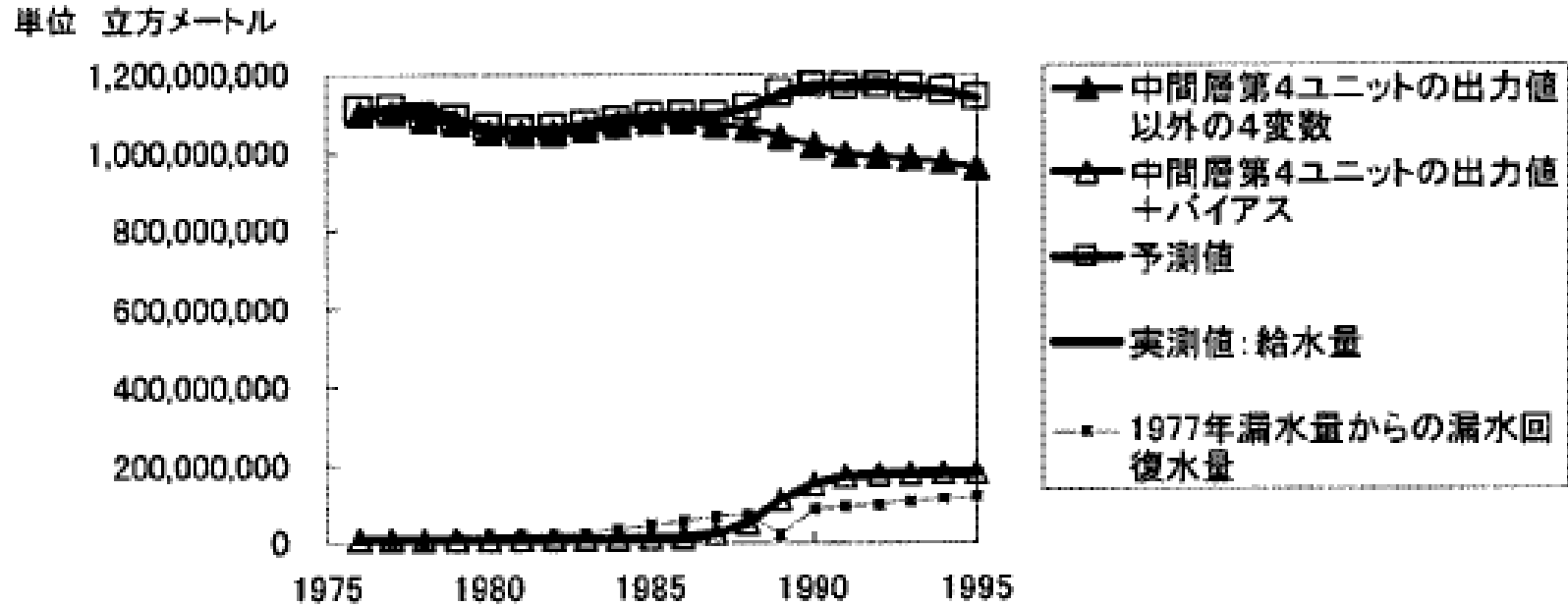
なぜ機械はデータで賢くなるのかを認識すべき

○任意関数自動近似技法

- 多項式型回帰（応答局面法）⇒ニューラルネットワーク系統計的機械学習
- Huberの射影追跡 + 不連続分布と連続分布同時近似可能な非線形関数
 - ボルツマンマシン：温度パラメータが高ければステップ関数（異質性自動分類）も近似可能
- 変数間交互作用の消去と非線形主効果（層別）の合わせ技で任意関数近似： $4XY = (X+Y)^2 - (X-Y)^2$
- 古典的データ分析戦略の進化をニューラルネット系AIは概ね反映：ただしブラックボックス化
 - 特性値を上手に選択せよ：因子モデル，PLS⇒合成指標の最適探索
 - 層別を上手に行え：潜在クラスモデル⇒層別境界の自動探索

○近似関数の単純化（パラメータ節約）

- 高次因子モデル⇒深層学習：階層化による単純構造での近似
 - 合成指標による予測方式自体の階層的クラスター表現：実はパラメータ数節約している（そうは思われていない）



- 第4ユニット
- 温度最大
- 時代変化表現

図 2. 東京都 23 区給水量と予測値と回帰係数 × 中間層第 4 ユニット出力値以外の説明変数の和、回帰係数 × 中間層第 4 ユニット出力値 + 定数、漏水回復水量グラフ

人工知能のある中間層は、
1980年代後半に水道利用量計測で起きていたことを予測損失最小化するために
自動的に学習

浅野・椿(2002): 応用統計学会誌
3層ニューラルネット回帰可視化 (シグモイド分解)

正則化あるいはベイズモデル戦略

○モデル選択→
正則化＝
ベイズモデル事前分布の
最適化

- ・ 許容性のあるベイズ決定の中で
パフォーマンス最適化

○逆問題（制御問題）の
正則化への注意

- ・ 非予測トレーニングデータの
平均・分散を予測に用いるのか？
- ・ 田口T法, ナイーブベイズを見直すべき

予測値を束ねる戦略

○複数単純予測値の情報量按分最適結合

- 個人の知より集団の知
- **Bartlettの因子得点推定, 田口玄一(品質工学)のT法が基本原理**
 - ➡Bugging, Random Forest

○予測誤差を改善する予測ロジックの追加戦術

- **回帰分析改善のPDCAサイクルを回している**
 - **回帰診断や田口玄一の実験的回帰分析➡Bugging**

（独）統計センター時計技術研究課
家計調査消費項目600分類への活用想定
機械学習研究

例：単純な 産業21分類教師付き学習

2値特徴量：(財務変数＋層別基準)を
そろえて1996年上場企業産業分類

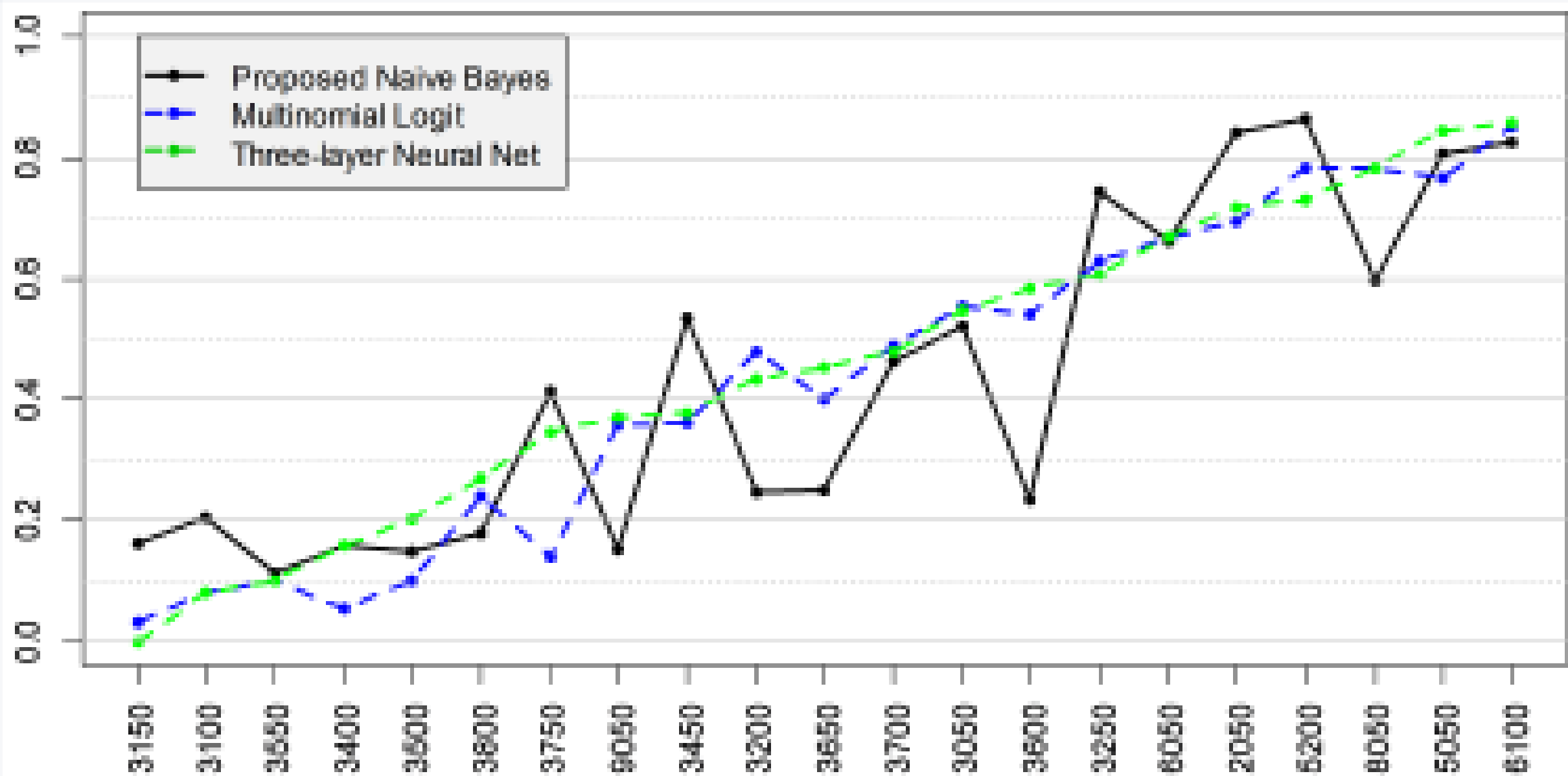
仮説：産業分類が財務情報に影響
逆問題：T法を離散データに拡張
ミニマックス的・ナイーブベイズ

単純な順問題型予測
多項ロジットモデル
3層ニューラルネット

Table 1 Industry classes for the dataset classification

Industry Class	Securities Classification Code	Number of Companies
Construction	2050	177
Foods	3050	119
Textiles and Apparels	3100	98
Pulp and Paper	3150	31
Chemicals	3200	157
Pharmaceutical	3250	43
Glass and Ceramic	3400	57
Iron and Steel	3450	58
Non-ferrous Metals	3500	40
Metal Products	3550	70
Machinery	3600	192
Electric Appliances	3650	203
Transportation Equipment	3700	103
Precision Instruments	3750	29
Other Products	3800	67
Land Transportation	5050	51
Warehousing and Harbor Transportation	5200	37
Whole Sale Trade	6050	181
Retail Sale Trade	6100	149
Services	8050	37

Correct classification rates of the classifiers



計算機環境

☆パフォーマンス
最適予測探索に
資する計算機パワー

モデル選択→価値関数
クロスバリデーション
評価などの
自動最適化

その他のアイデア

超高次元内での分類・
回帰の**効率化**
・カーネルトリック
(Support Vector Machine)

階層性を意識したデータサイエンス教育

初級者用

- 入力制約が少ない,
- 出力システム理解可
- 中程度の予測精度
- Classification and Regression Tree :
 - 自動層別技法
 - 品質管理の基本

中級者用

- 入力に加工を行い,
- 出力システム理解可
- 中程度の予測精度
 - 交互作用や非線形性を考慮
 - セミパラメトリック回帰モデル
- 散布図に線を引く
- 通常の回帰モデル
 - 散布図に式を適合
 - 射影追跡回帰

上級者用

- メカニズム解釈を含むモデルで高精度予測
- 技術者が統計家との協働で作る非線形パラメトリック予測モデル
- 一般化可能
- 技術知識の獲得

ベンチマーク用 ブラックボックス

- 入力制約が少ない, システムは分からないが, ベストパフォーマンス
- Random Forest
- 深層学習
 - ブラックボックス
 - チューニング大変

ご清聴ありがとうございました

データ・サイエンス時代でも人間はより賢くなるものと信じます
そのためにも良いデータ分析事例を社会共有することが大切です

日科技研のその種の活動が発展することを期待します。

別件ですが政府統計へもご協力よろしくお願い申し上げます

本著作物は原著作者の許可を得て，株式会社日本科学技術研修所（以下弊社）が掲載しています．本著作物の著作権については，制作した原著作者に帰属します．

原著作者および弊社の許可なく営利・非営利・イントラネットを問わず，本著作物の複製・転用・販売等を禁止します．

所属および役職等は，公開当時のものです．

■公開資料ページ

弊社ウェブページで各種資料をご覧ください <http://www.i-juse.co.jp/statistics/jirei/>

■お問い合わせ先

(株)日科技研 数理事業部 パッケージサポート係 <http://www.i-juse.co.jp/statistics/support/contact.html>