人文科学とコンピュータ　32－9
（１９９６．　１１．　１５）

「宗門改帳」を入力史料とした古文書画像データベースの構築

川口　洋、上原邦彦

帝塚山大学経済学部
〒６３１　奈良県奈良市帝塚山７－１－１

本稿では、江戸時代における人口分析システムを開発するための第一段階として、「宗門改帳」古文書画像データベースの構築した。「宗門改帳」（しゅうもんあらためちょう）とは、17世紀末から19世紀中期の期間に、原則として集落単位に毎年作成されていた人口史料の総称である。従来の研究方法では十分保障されていなかった、史料読解から文字データ入力に至る研究過程の再現性を、古文書画像データと文字データの両者を同一画面上で検索・表示することによって確保した。さらに、史料読解から文字データ入力までの作業過程を短縮するために、年齢を表記した漢数字を対象として、古文書文字の自動認識に関する実験を行った。

IMAGE DATABASE FOR ANALYZING THE JAPANESE RELIGIOUS INVESTIGATION REGISTERS

Hiroshi KAWAGUCHI and Kunihiko UEHARA

Faculty of Economics, Tezukayama University
7-1-1 Tezukayama, Nara, 631, Japan
kawag@tezukayama-u.ac.jp
uehara@tezukayama-u.ac.jp

We have constructed the image database for analyzing the Japanese religious investigation register so called "Shumon-Aratame-Cho(SAC)". This database is planned in order to make the process of outputting the demographic statistics from the SAC data easier and faster, to guarantee the quality of the process, to preserve the present condition of the SAC data and to share the sourse data with historical demographers. We also experimented the character recognition on age data which are expressed in handwritten old Chinese figures.

1. はじめに

本研究では、17 世紀末から 19 世紀中期の期間に、原則として集落単位に毎年作成されていた「宗門改帳」と総称される古文書を源史料として、江戸時代における人口分析をできるかぎり自動化するシステムの構築を目指している。本システム構築の第一の目的は、「宗門改帳」から人口学的指標を、正確、迅速に算出することによって、研究を支援する点にある。さらに、「宗門改帳」を読解した文字データ、「宗門改帳」の古文書画像、および人口学的指標算出プログラムを、システム構築に直接携わらない一般研究者に公開することによって、人口分析の研究過程を再現し、利用者の批判をデータベース改良にフィードバックできるシステムの構築を目指している。

「宗門改帳」から人口学的指標を算出することを目的としたシステムの先行研究として、川口(1990、1992、1995)、小野(1993)、木下(1996)が上げられる。三者ともに、作業時間の短縮という当初の目的は達成しているが、文字データを基礎としてシステムを構築しているため、「宗門改帳」読解から文字データ入力に至る研究過程の再現性が十分確保されているとはいえない。

本システム設計の基本方針は、以下の 4 点である。

① 「宗門改帳」の読解から人口学的指標算出に至る研究過程の作業量、作業時間を短縮する。

② 「宗門改帳」の読解から人口学的指標算出に至る研究過程の再現性を確保する。

③ 「宗門改帳」の現況を画像情報として保存する。

④ 「宗門改帳」の史料情報、基本的な人口分析の方法を、研究者間で共有する。

本稿では、人口分析システム開発の第一段階として、「宗門改帳」古文書画像データベースを構築した。システムの概要は、図 1 に要約される。

2. 入力史料の概要

陸奥国会津郡小松川村他五ヶ村(現在の福島県南会津郡下郷町大字小松川)の「宗門家別人別改書上帳」を古文書画像データベース構築のための試行資料として選択した。寛政 4 (1792) 年から慶応 2 (1866) 年) までの 75 年間については、連続して分析することができる。この期間のうち、寛政 11 (1799)・13 (1801)、文化元 (1804)・7 (1810)、天保 5 (1834)・12 (1841)、安政 3 (1856)、万延元 (1860) 年については、史料が散逸しているが、前後の年次の史料から内容を復原することができる。

この 75 年間における「宗門家別人別改書上帳」の書式は、図 2 のようにほぼ固定している。住居と家計を共にする世帯に相当する記載単位ごとに、旦那寺の所在地・本末関係・宗派・名称、持高、家屋規模、屋根の材料、構成員の名前、筆頭者との続き柄、年齢、異動、牛馬数、世帯規模が記録されている。

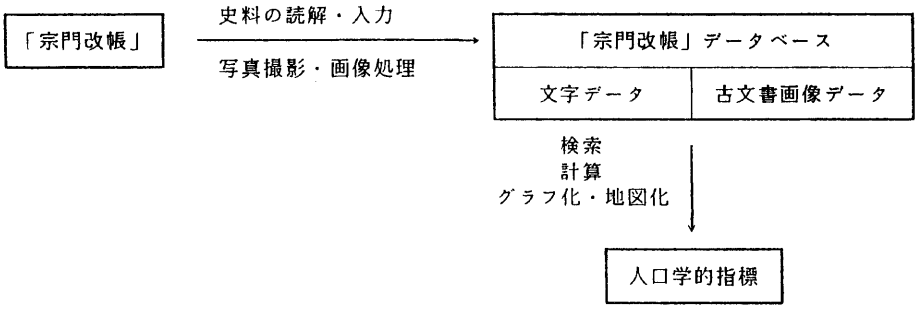
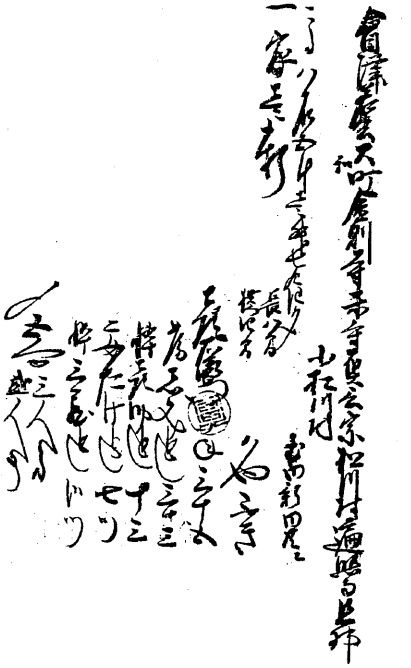


図 1　人口分析システムの概要



會津若松大和町金剛寺末寺真言宗松川村遍照寺旦那
小松川村
高　八石五斗壱升七合四夕　本田新田共ニ
一、家壱軒　長　八間　かやふき
横　四間
喜右衛門　印　年　三十五
女房　しめ　とし　三十三
倅　喜助　とし　十三
二女　たけ　とし　七ツ
倅　三蔵　とし　弐ツ
〆　五人内　三人男
弐人女

図2　入力史料の書式

佐藤仁夫家所蔵「文久三年　宗門家別人別改書上帳　小松川村、沢入村、寺山村、寺村」

3. 古文書画像の入力

鮮明な古文書画像を表示できる点、史料の現地調査に携行する機器が簡便である点、現在までに蓄積された人口史料のフィルムを継承することができる点などを考慮して、「宗門家別人別改書上帳」を写真撮影した後、PHOTO CD に書き込み、保存した。写真は、照明台の上にカメラ（NIKON F2、NIKOR 55mm/F3.5）を固定して、FUJICOLOR SUPER G ACE 400 のフィルムで、史料の見開き２頁を１画像として撮影した（川口他、1996）。 PAINT SHOP を用いて、PHOTO CD から 1536*1024 DOTS の解像度で読み取り、１世帯を１画像に編集した後、 TIF 形式で保存した。

4.「宗門改帳」古文書画像データベースの概要

MICROSOFT 社の ACCESS を DBMS として、「宗門改帳」古文書画像データベースを構築した。同データベースは、ア）個人情報、イ）世帯情報、ウ）史料書誌情報、エ）古文書画像情報の４テーブルから構成されている。ア）からウ）は文字データ、エ）は文字データと画像データで構成される。各テーブルのフィールドは、以下に示される。

ア）個人情報テーブル

集落名（国郡村）、西暦、世帯番号、個人番号、名前（ローマ字）、名前（漢字）、年齢、性別、筆頭者との続き柄、配偶関係、宗教・宗派、旦那寺の所在地、旦那寺、異動事項、異動内容、村役人

イ）世帯情報テーブル

集落名（国郡村）、西暦、世帯番号、筆頭者名（ローマ字）、筆頭者名（漢字）、家族人数（男性）、家族人数（女性）、譜代下男人数、譜代下女人数、質券下男人数、質券下女人数、同家人人数（男性）、同家人人数（女性）、世帯規模、

世帯構造コード、同居世代数、持高（石）、牛数、馬数、家屋規模（縦＊横、間）、屋根材料、

ウ）史料書誌情報テーブル

集落名（国郡村）、西暦、史料作成年月日（和暦）、史料名、史料作成者名、所蔵者名

エ）古文書画像情報テーブル

集落名、西暦、世帯番号（または、表紙、または、末尾）、古文書画像

集落名、西暦、世帯番号を検索語として世帯単位にファイルに保存されている古文書画像を表示することにより、ア）からウ）の文字データとエ）古文書画像データを、同一画面上で対照することができる。したがって、古文書文字に誤読の疑いを抱いた場合、あるいは、原史料の文字の配列、筆跡、印形といった文字データとして登録することのできない画像情報を参照したい場合などにも、データベース利用者が、データベース構築者の研究過程を再現、批判できる環境を整えることができた。

5. 古文書文字の自動認識

「宗門改帳」古文書画像データベースを構築する場合、最も長い作業時間を必要とするのは、史料の読解、入力である。この作業過程を自動化できれば、「宗門改帳」読解から人口学的指標算出までの作業時間を、飛躍的に短縮することができる。古文書文字の自動認識への第一段階として、年齢、牛馬数、世帯規模、家屋規模、持高などに用いられる漢数字を対象として、実験を行った。漢数字で表記される情報の中で、年齢は、結婚年齢、出産年齢、死亡年齢、夫妻の年齢差など、民衆の生活を復原するうえで重要な人口学的指標を算出する場合、とくに重要な基礎的データとなる。年齢に記録される漢数字の種類は、限定されるうえに、古文書のレイアウトなかでも、図2のように、世帯構成員の名前の下のほぼ固定した位置に記録されている。そのためまず、年齢の漢数字について自動認識を試みる。

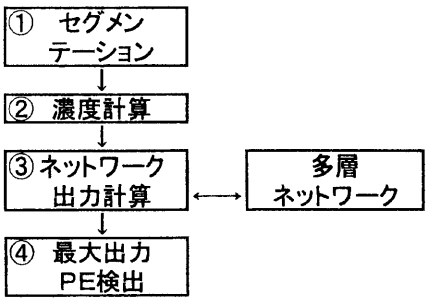


図3. 文字認識のブロック図

ここでは、全般的な方法として、ニューラルネットを用いた手書き文字認識と同様の方法を採用する（小川、1994）。この方法を選択した理由は、特徴抽出による識別処理（山田、1995）に比べ、漢数字のすべての特徴を取り入れられると期待するからである。さらに、顕著な特徴を入力プロセッシングエレメント（ＰＥ）に採用することも可能である。この古文書文字認識の処理の流れは、図3のように、セグメンテーション・濃度計算・ネットワーク出力計算（3層ニューラルネット）・最大出力ＰＥ検出となる。認識過程のおのおのの処理は、以下に要約できる。

① セグメンテーション：レイアウト解析（劣化画像復元を含む）、行切出、文字切出

② 濃度計算：標本点選択、マスクパターン決定、濃度特徴計算

③ ネットワーク出力計算：学習済3層ニューラルネットによる出力ＰＥ計算

④ 最大出力ＰＥ検出：最大出力ＰＥ検出、対応文字決定

セグメンテーションについては先行研究（柴山他、1996a、1996b、山田、1996）を活用することとして、本稿では濃度計算からはじめる。ニューラルネットの学習過程では、濃度値を計算後、その濃度値を入力層からの出力としてバックプロパゲーション法で学習する。これを3層ニュー

ラルネットの範疇で実行し、結合係数を出力層各ＰＥの教師付き学習により求める。バックプロパゲーションにおいての結合係数の最適化法としては、誤差（エネルギー）関数の最小値問題としてシミュレーテッドアニーリング法（S. Kirk Patrick 他、1983）を応用する。この方法を使う利点は、ある条件を満たしながら温度を下げるスケジュールをとると、必ず誤差（エネルギー）関数の最小値に収束する点にある。ただし、この方法は計算時間を要するので何らかの高速化、もしくは別の方法を組み合わせる必要がある。

今回提案する高速化法は、巡回セールスマン問題（ＴＳＰ）に適応されている実空間繰り込み群的アプローチ（宇佐美・加納、1995）である。これは、ＴＳＰ自身が都市数に関してスケール普遍であることを利用している。具体的には、都市を囲む正方形の枠を4^n等分する。そのとき、あるセルに都市が含まれるならそれらの都市を代表点で置き換えて、各々のｎとｎ＋１回目の関係が存在するとしてセルを細分していく方法である。

本稿の問題でいえば、入力層ＰＥ数は標本点の数によるが、結合係数の最適化計算時間は、標本点の数に依存する。そこでこの標本数に関して実空間繰り込み群的アプローチを使用するものである。認識文字種が多くなる時には、このような高速化を考えておく必要がある。この高速化法の有効性は、ＴＳＰのスケール普遍性がこのアプローチの有効性と関係していると考えられていることから、文字認識の標本点についても有効と考えられる。

6. おわりに

本稿では、江戸時代における人口分析システムを開発するための第一段階として、「宗門改帳」古文書画像データベースの構築した。従来の研究方法では、十分保障されていなかった「宗門改帳」読解から文字データ入力に至る研究過程の再現性を、古文書画像データと文字データの両者を同一画面上で検索・表示することによって確保した。さらに、史料読解から文字データ入力までの作業過程を短縮するために、年齢を表記した漢数字を対象として、古文書文字の自動認識に関する実験を行った。

江戸時代における人口分析システムを構築するための課題として、１）「宗門改帳」古文書画像データベースから人口学的指標を算出するプログラムを作成する、２）インターネットを通じてシステムを公開する、３）分散処理の準備を行う、といった点があげられる。

古文書文字の自動認識については、研究をようやく着手した段階にあり、劣化画像の修復、文字の切出、漢数字以外の文字認識といった課題が山積している。

付記）　本報告は、文部省科学研究費重点領域研究「人文科学とコンピュータ」公募研究、平成７年（課題番号：07207237）、平成８年（課題番号：08207231）の補助を得て行った研究成果の一部である。

参考文献

１）宇佐美義之、加納義樹（1995）人工知能学会全国大会論文集、９、p.377

２）小川英光編著（1994）『パターン認識・理解の新たな展開』、電子情報通信学会、pp.26-41

３）小野芳彦（1993）文化系の計算機利用 II　－データ入力のユーザーインタフェース（歴史人口学の場合）－、日本研究、８号、pp.165-182

４）川口　洋（1990）江戸時代における人口分析の方法　－奥会津地域における「宗門改人別家別帳」のデータベース化を事例として－、歴史地理学、151、pp.16-33

５）川口　洋（1992）「宗門改帳」データベース・システム（DANJURO）の改良、情報処理学研究報告、vol.92 no.19、pp.1-8

６）川口　洋（1995）コンピュータを用いた江戸時代における人口分析の方法、人文学と情報処理、７号、pp.54-58

７）川口　洋・上原邦彦・上島紳一（1996）江戸時代における人口分析システム開発に関する予

備的研究、平成７年度科学研究費（重点領域研究）研究成果報告書『人文科学とコンピュータ　1995年』、pp.501-508
８）木下太志（1996）歴史人口学における人口指標算出およびグラフィック化のためのプログラム開発、平成７年度科学研究費（重点領域研究）研究成果報告書『人文科学とコンピュータ　1995年』、pp.529-542
９）芝山　守・富田浩章・西門秀人・荒木義彦（1996a）ビデオによる古文書画像の入力と文字抽出について、平成７年度科学研究費補助金（試験研究Ｂ－１）研究成果報告書『東洋学研究における大量マルチメディア情報の提供方式の研究と実用化』、pp.45-53
10）S. Kirk Patrick, C.D. Gelatt.Jr. and M.P. Vecchi, 1983, "Optimization by Simulated Annealing", Science, 220, pp.671-680
11）富田浩章・柴山　守・荒木義彦（1996b）２値化レベル制御による古文書画像の文字セグメンテーションとパターン字書について、情報処理学会研究報告、vol.96 no.42、pp.7-12
12）山田奨治（1995）高次局所自己相関特徴による古文書かな文字認識、情報処理学会研究報告、vol.95 no.14、pp.21-30
13）山田奨治（1996）人文科学研究のためのパソコン画像処理システムのインターフェース、情報処理学会研究報告、vol.96 no.73、pp.1-6