Computational Archives of Population Dynamics and Migration Networks as a Gateway to Get Deep Insights into Hard-to-reach Populations: Research on Taiwan Indigenous Peoples

Ji-Ping Lin
Academia Sinica
Taipei, Taiwan
E-mail: jplin@sinica.edu.tw

Abstract—This paper highlights research on constructing big computational archives of hard-to-reach populations (HRPs), using Taiwan Indigenous Peoples (TIPs) as an example. The research uses archives of (1) anonymous individual-level migration flows computed from population dynamics data and (2) Taiwan indigenous community data (TICD) to illustrate characteristics of HRPs which were unknown before. The research suggests that computational HRP networks (e.g., migration networks) help overcome barriers to accessing HRPs and promote mutual understanding. The archives of Taiwan Indigenous Peoples Open Research Data (TIPD) are a research data source, with archives of address geocoding, population dynamics, and indigenous communities being most relevant to TIPs network systems. The migration flows are computed at the individual level and have unveiled various dimensions of HRP networks that were invisible before. The newly computed TICD archives enable us to trace migration flows of TIPs within and between indigenous communities and urban localities at the individual level in the context of ethnic lineages. The research findings suggest that strengthening intra- and inter-ethnic network connections serves as an effective measure to get deep insights into HRPs.

Keywords—hard-to-reach population; migration network; population dynamics; ethnic lineage; TICD; TIPD

I. INTRODUCTION

Hard-to-reach populations (HRPs hereafter) are those hard to access due to geographical location or social status. They are characterized by being vulnerable, excluded, and hidden in a society. The ability to access HRPs enables us to build insights into various issues they encounter and thus help us to design effective policy measures. Recruiting hard-to-reach populations has long been a big challenge for study. There are various barriers relevant to recruiting hard-to-reach populations. The most crucial challenges are: how to (1) label the population for study; (2) overcome mistrust of participants when reaching HRPs; (3) overcome legal and ethical issues to protect HRPs privacy and confidentiality [1][2].

This paper highlights the author’s research on constructing HRP big computational archives, using the population of Taiwan Indigenous Peoples (TIPs hereafter) as examples. Taiwan Indigenous Peoples (TIPs) are one or more subgroups of the Austronesian family. Although the population size of TIPs is not big, TIPs are known in the complexity of ethnic relations, socioeconomic organization, cultural plurality, political structure, and linguistic diversity, etc. Understanding such complexity has long been a big challenge. One main reason lies in the lack of data on TIPs before the year of 2000.

In spite of a rich body of ethnographic data, mostly recorded in text and image format, contemporary and systematically collected detailed numerical individual-level datasets are not available until 2010. This offers us a unique opportunity to study contemporary TIPs. In the past decade, the author devotes to contemporary TIPs study. One important outcome is the construction of a big archival numerical datasets termed as “Taiwan Indigenous Peoples Open Research Data” (TIPD hereafter, see https://osf.io/e4rvz). TIPD archives serve as the main data source for research [3][4].

In this paper, the author mainly uses three broad categories of computational archives to illustrate main characteristics of HRPs. These include: (1) computational archives of population dynamics in TIPD; (2) new computational archives of anonymous individual-level migration flows computed from TIPD’s archives of population dynamics; and (3) computational archives of indigenous community data. Archives of indigenous communities serve as access nodes to HRPs, while anonymous individual-level migration flows of indigenous peoples serve as the level of kin- and friendship network linkages within and among indigenous and non-indigenous communities.

Strictly complying with legal and ethical requirements [5][6], the research uses household registration data of TIPs to help label potential hard-to-reach populations. The research aims to demonstrate (1) that the computational big archives that are computed based on computational archival science shed lights on various characteristics of HRPs not known before; and (2) that they enable us to utilize HRP networks (e.g., migration networks) to...
overcome invisible barriers of accessing HRPs and pave a path to promote mutual understanding and trust [2].

II. THEORETICAL FOUNDATION AND COMPUTATIONAL INFRASTRUCTURE

A. Theoretical Foundation of Migration Flows as Networks

The development toward diversity in various aspects in the past three decades has accelerated migration of TIPs, mostly rural-to-urban migration [7]. Similar to other disadvantaged population in other countries, TIPs are characterized by the following features: having higher fertility/mortality, shorter life expectancy, higher migration likelihood, less individual income gains, and less social mobility, etc. [8]

From the perspectives of various schools of migration theory, migration has the effects of promoting both individual lifetime income gains and social mobility [9]. For example, the neo-classical school of migration theory suggests that migrants are expected to have much higher net lifetime income gains than non-migrants and that migration has the effect of promoting individual wellbeing and social mobility. The school of “new” economics of migration regards migration as triggered by relative deprivation and holds that migration eventually reduces inequality in various aspects like income, education, work opportunity, health, etc. [10].

The most noteworthy feature of the aforementioned characteristics associated with TIPs is that TIPs are much more migratory than other people. If migration theory is correct, the very migratory TIPs are expected to have benefitted from the process of migration, and the gap in, e.g., incomes and inequality between TIPs and non-TIPs should have converged gradually. Unfortunately, this is not the case [8].

This is unlikely to be the flaws of the existing migration theoretical framework. Based on the experience of fieldworks, the author gradually realizes that this paradox might be an outcome of the complexity and particularity of the ethnic relationship: the degree of connection between the internal ethnic groups and the lack of strength of the social network connections with the non-indigenous people. As a result, exploring the individual-level migration flows that reflect personal kinship and friendship networks by each ethnic group is an effective way to get insights into the aforementioned network contradictions.

B. Digital Infrastructure and High Performance Computing

As indicated by Lin [4], computing and enriching TIPD archives are not simple sequential data processing. It adopts distributed computing [11][12], multi-threading, and in-memory computing methods to enhance both computing performance and efficiency [13]. Because the amount of accumulated source datasets that are used to compute TIPD archives in 2018-2021 increases three times the size of data used in 2013-2017, computing load grows substantially and hardware infrastructure has been upgraded to meet the computing requirement.

In order to optimize and maximize both the performance and efficiency of archival computing processes, digital facilities are assembled and fine-tuned by the author in the lab. As shown by Figure 1, assembling and fine-tuning digital infrastructure are a routine task and have taken the author a lot of time. The main hardware upgrades are: (1) a dual Intel Xeon 2683v4 CPUs 32-core workstation with 768GB DDR3-2600 DRAM, and (2) an AMD 64-core 3990X CPU workstation with 256GB DDR4-3600 DRAM and two high-speed 4TB and 8TB CPU-direct-access PCIe-G.4 NVMe RAID0 storages. The main software tools used in computing are Embarcadero RAD Studio 10.1 Berlin edition and SAS v9.4.

![CrystalDiskMark 8.0.4 x64](image)

**Figure 1.** Assembling and fine-tuning computing digital infrastructure

The author would like to point out that successful assembling, upgrading and fine-tuning computing systems are very crucial to archival computing tasks. Digital infrastructure and hardware settings play a crucial role during the computing processes of constructing TIPD archives. The main computing method is in-memory computing, i.e., the integration of accelerating CPUs and DRAM speed and broadening I/O buses data transfer speed (see Figure 1). It is made possible through the processes of
(1) fine tuning BIOS settings and (2) making full use of DRAM and RAID0 NVMe SSDs.

The programming codes are designed for in-memory computing purposes in the sense that all computing tasks are implemented in the computer’s internal memory, with CPUs and internal memory being overclocked, and the data transfer rate I/O bus between CPUs and DRAM being accelerated. In the past three years, high-speed and high-capacity CPU-direct-access storage devices, e.g., NVMe RAID0 storage via PCIe G.4, are impressive. They not only help overcome traditional I/O bottleneck of storage devices, but also facilitate to increase in-memory computing performance substantially. For example, see Figure 1.

III. DATA AND METHODOLOGY

A. Data

TIPD are research data source, consisting of digital archives from 2007 to 2021. For detailed information about source data sets that are used to compute TIPD archives, see [4]. The research program of constructing TIPD archives begins in 2013. With the support of funding agencies, the relevant work of extending and updating the content of TIPD archives remains in steady progress in 2021. Not only are knowledge and implementation ability of computational archival science required, practices of the research are also required to be in accordance with the real world domain knowledge. Consequently, the author also has conducted a series of fieldwork surveys and expert interviews to verify the validity of computational archives. In short, expertise in understanding the real world enables us to enrich and extend the content of computational archives and to avoid potential risk of departing from conventional wisdom.

TIPD includes the following broad categories of computational archives: (1) categorical data, (2) multi-dimensional data, (3) population dynamics data, (4) household structure data, (5) indigenous community data, and (6) data visualization. Methodologically based on data science, the construction of TIPD also fully complies with open science principles and open data requirements [14][15][16]. It is available on Open Science Framework (OSF) at https://osf.io/e4rvz/. For details about the design of computational infrastructure, data engineering, computation methods, categories of open data, and data sharing in TIPD, please refer to [4].

By June of 2021, the total TIPD archives amount to more than 7,000 digital files and 210GB in size. The development of TIPD archives includes two different strategies from 2013 to 2021. In the period from 2013 to 2017, the development strategy of TIPD mainly focuses on (1) establishing a robust digital infrastructure and an automated computing system that enable us to integrate and enrich various data sources; (2) developing computing methods that transform confidential data into open data; (3) building up an open data repository system to promote research collaboration and transparency [4].

In the period from 2018 to 2021, the main strategy of developing TIPD archives has moved to improving quality and increasing diversity of digital data archives. One important advance is the development of an automated address geocoding computing system. It allows us not only to parse precise geo-information for big digital records in batch mode, but also enables us to get rid of various restrictions from Google Map APIs.

Because source data archives are based on monthly collection, the latest recomputed TIPD archives provide us with much more detailed information on temporal and spatial dynamics. For example, TIPD’s new population dynamics archives allow us to explore very detailed birth dynamics, death dynamics, and migration dynamics from both spatial and temporal perspectives. Moreover, TIPD’s new archives about indigenous communities offer us detailed insights into the evolution of indigenous communities over time and space, including population structure, ethnic relations, interactions with non-indigenous population, etc. In short, the archives in TIPD most relevant to TIPs network system are archives of population dynamics and indigenous communities.

B. New Method of Improving Address Geocoding

Since the beginning of this research program in 2013, one important goal of the research is to collect individual geographical location information through household address geocoding method. The research at the initial stage used Taiwan Address Matching System (TAMS) to locate indigenous peoples’ households. However, the research found that TAMS can only locate 85% of total population. The remaining 15% population that TAMS fails to locate are mostly located in Central Mountain Areas and eastern Taiwan. It is worthy of stressing that the areas that TAMS fails to locate individuals are main residential areas of HRPs.

When the number of accumulated individual archives increased rapidly as the research program proceeded, the author realized that TAMS had to be abandoned and that there was a need to seek an alternative address geocoding system. Google geocoding system provides us with more precise and accurate geocoding information. More importantly, it enables us to locate HRPs in Taiwan’s Central Mountain Areas and eastern Taiwan. The research thus moved geocoding system from TAMS to Google’s geocoding system in late 2016.

However, Google system imposes many restrictions in address geocoding, e.g., limitation on the number of addresses which can be geocoded. To overcome this constraint, the author develops a method in late 2018 that
allows us to make voluminous address geocoding by parsing Google Map information in automated batch mode. The author continues to improve this new geocoding method and it now works smoothly without any problem.

As a result, the research finally achieves to re-locate more than eighty million individual archives from 2007 to 2021, with each individual’s latitude and longitude information being recorded in WGS84 projection coordinate system. Using digital terrain model (DTM), the corresponding altitude information for each pair of latitude and longitude is also collected in address geocoding archive.

C. Protecting Privacy of Individual Spatial Information

Precise individual geolocation information may fail to protect individual privacy, even though personal information like gender, age, etc., is not revealed. To comply with legal and ethical requirements, the research uses “truncated” address rather than “full” address information to locate all individual geocodes. For example, an address that reads “128, Lane 16, Section 2, Paisan road, 27 Lin, Chongyen village, Nankang district, Taipei city, Taiwan 115” will be truncated as “Lane 16, Section 2, Paisan road, 27 Lin, Chongyen village, Nankang district, Taipei city, Taiwan 115”.

Moreover, the research further adopts “population center” of a given area as a proxy to represent an individual’s exact location information. The population center of a given area is calculated as follows: suppose a given area has n individuals, we calculate the mean value of n latitudes and that of n longitudes associated with the n individual in the given area. The mean latitude and mean longitude derived from the n individuals thus refers to the population center of the given area. Likewise, altitude information is transformed using the same method. In other words, individual location information of all persons in this given area are the same, i.e., the population center of the given area.

Geocoded individual location information is added to all archives while processing source datasets, but only population centers are available in TIPD archives. Each individual’s location information is represented by five categories of population centers that are associated with five different geographic units. The five categories of geographic unit consist of (1) four regions, (2) 25 prefectures/cities (equivalent to county in the U.S.), (3) 365 township units; (4) 8,700 village units, and (5) 576,800 “Lin” (鄰) units (equivalent to census block in the U.S.). Consequently, all individuals reside in the same area of a given geographic unit have the same location information. For example, location information for individuals in a given village is represented by the population center of the given village.

IV. TYPES OF COMPUTATIONAL HRPs ARCHIVES

A. Individual-level Geolocation Information

As indicated that all TIPD archives, including the geolocation information archive, for all individuals from 2007 to 2021 have been recomputed, with newly developed automatic address geocoding system being applies to compute new individual location information. Reconstructing archive of precise geolocation information and integrating it with other existing archives lay foundation for advanced TIPs network analysis (e.g. ethnic relation network, kinship network, friendship network, marriage network, migration network, and thus social mobility network, etc.) as well as advanced predictive analytics in the context of individual spatial interaction behavior over time.

Figure 2.a. Address geocoding archives and population distribution of all TIPs

Using new address geocoding archive, Figures 2.a & 2.b demonstrate the latest TIPs spatial distribution by all TIPs and selected ethnic TIPs groups out of all sixteen TIPs ethnic groups. The characteristics of TIPs population distributions are summarized as follows: (1) each ethnic group has its own distinct distribution pattern and scope; (2) TIPs mostly reside in periphery of metropolitan areas, Central Mountain Areas, and eastern Taiwan; (3) most TIPs choose to reside along the main streams of a drainage basin in rural and mountain areas.
As indicated, in addition to constructing archive of latitude and longitude information, the corresponding archive of altitude information is created. Figure 3 demonstrates TIPs distribution by altitude level. It indicates that most TIPs reside in low- and moderate-altitude places (below 1,200m and 1,200-2,400m, respectively), and only 1% of TIPs reside in high-altitude areas (above 2,400m).

B. Individual-level Migration Flows

Archives of population dynamics in TIPD are utilized to construct migration flows. All population dynamics datasets from 2013 to 2021 have been recomputed and serve as one of the most innovative outputs in the TIPD. For details about archives of population dynamics, please refer to “TIPD/6_PopulationDynamicsData" (https://osf.io/e4rvz/files/?view_only=8764e9e3d9f543eeb4bf507e21dfc6fa).

Population dynamics refer to three types of dynamics, i.e., dynamics of the birth process, dynamics of the death process, and dynamics of the migration process. The data model for constructing population dynamics is straightforward. It is mainly based on (1) comparison and (2) record linkage of two population data sets at time point 1 and time point 2 [17][18].

In terms of comparing two population data sets, the population members found in the data at time point 1 but who are not available in the data at time point 2 are termed the “decreased population” in this research; likewise, the people who are not available in the data at time point 1 but become available at time point 2 are termed “increased population”. The people who remain in the period of both time points 1 and 2 are termed the “intact population”.

Record linkage is applied to incorporate information from both time points 1 and 2 for the “intact population”. The “intact population” is dichotomized into two subgroups of population: the “stay-put” group (i.e., those who don’t make migration between time points 1 and 2) and the group who make internal migration (i.e., those who make migration at time points 1 and 2 in the period).

The author uses Figure 4 to illustrates birth, death, and migration processes of TIPs derived from population dynamics archives. Based on archives of population dynamics, population change of TIPs is mainly shaped by birth and death processes, and international migration is negligible. Areas with higher poverty but with less inequality have higher incidences of birth and death, while areas with much lower poverty but having high inequality tend to have higher level of birth and lower incidences of death. Moreover, the research finds that net population growth rate of TIPs from 2013 to 2021 is about 6.8%, with natural increase rate and social increase rate being 3.2%
and 3.6%, respectively. Meanwhile, the rate of intact population who made internal migration is about 26.6%.

Figure 4. Archives of TIPs population dynamics: the birth, death, and migration processes

Figure 5. Archives of TIPs migration flows: all TIPs and selected ethnic groups

Each individual has been located, with personal location information on latitude and longitude being labelled in individual record. The research uses individual migrants in “intact population” archives to construct individual flows of migration through the way of drawing a line from the origin to the destination of migration. Figure 5 illustrates selected migration flows at individual level by selected TIPs ethnic groups. As suggested by Figure 5, migration of TIPs is not only a highly selective process, but each ethnic TIPs has its own distinct pattern of migration. It is worth noting that if migration flows of TIPs are viewed as ethnic networks, we find that the connections of TIPs inter-ethnic networks are characterized by weak connections.

Archives of population dynamics from 2013 to 2021 not only enables us to explore ordinary migration flows, but also allows us to examine different types of migrations flows, including primary migration, return migration, and onward migration. Primary migration refers to the migration of those who did not make migration before. Return migration and onward migration are repeat migration of migrants. Return migration refers to the migration of migrants who choose to return to the origin of previous migration, while onward migration to the migration of migrants who migrate to other destinations. The three types of migration further suggest that various dimensions of HRP networks in terms of migration have been unveiled. Figure 6 demonstrates individual migration flows with respect to primary migration, return migration, and onward migration.

For the sake of protecting privacy, it is worth noting that the village-level population centers are applied to visualize individual-level migration flows between the origin and the destination. Detailed migration visualization for each TIPs ethnic group can be downloaded at:

- “TIPD/2_DataVisualization/1_TIPS_Migration” (https://osf.io/e4rvz/files/?view_only=8764e9e3d9f543eeb4bf507e21dfe6fa).
- Interactive visualization of migration flows is available at “TIPD/2_DataVisualization/3_Interactive Viz_Migration By Ethnic Groups” (https://osf.io/e4rvz/files/?view_only=8764e9e3d9f543eeb4bf507e21dfe6fa).

Figure 6. Migration dynamics of TIPs: primary, return, and onward migrations

C. Taiwan Indigenous Community Data (TICD)

There has been a rich body of historical archives, mostly ethnographic text data, about Taiwan indigenous communities since the early 17th century during the Dutch colonial period. However, traditional Taiwan indigenous communities remained not well explored until Taiwan became part of Japanese Empire in 1895. In particular,
high-land indigenous communities were the least known ones until the year of 1935 [19]. Unfortunately, we still have very limited knowledge about Taiwan indigenous communities and thus indigenous peoples, because detailed statistical archives and numerical data were not available until 2010.

Contemporary Taiwan indigenous communities have undergone tremendous changes, but their organizational infrastructure in terms of ethnical lineage system remains similar to that in early 1930s. In early 2017, the author has computed a tentative numerical archive of indigenous communities for evaluation purposes, but the outcome is not satisfactory. Because the calls for detailed aggregate statistical figures on indigenous communities from academicians and policy makers are very high, to construct a more comprehensive numerical archive and statistical figures of contemporary Taiwan indigenous communities becomes highly necessary. The author formerly began the computing research work of Taiwan indigenous community data in early 2018.

As demonstrated by Figure 8, TICD archive offers selected crucial statistic figures and numerical data for all 716 indigenous communities in Taiwan. The selected statistic figures include the following information: ethnical lineage, population size, sex ratio, educational composition, age composition, marital status composition, child dependency ratio, old-age dependency ratio, dependency ratio, population composition by ethnic group, population center. TICD also integrates descriptive text data about TIPs communities from other sources of historical archives.

Figure 7. Taiwan indigenous communities by TIPs ethnic groups

Based on official classification of indigenous community defined by Taiwan’s Council of Indigenous Peoples, the author makes uses of: (1) address information in source data to define the scope of each indigenous community; (2) an address geocoding archive to calculate the population center for each indigenous community; and

(3) source data to construct standardized statistical figures for each indigenous community. The author spends more than three and a half years to construct TICD archive, as shown in Figure 7. In late August of 2021, TICD archive was open to the public. For details about TICD archive, please refer to https://www.google.com/maps/d/viewer?mid=1M6FE6vKd212udITsvWnmFsaKavB_3TpF7&ll=24.092205710969726%2C120.89717315782033&z=6.
V. CONCLUDING REMARKS AND DISCUSSION

To gain more insight into the HRP, scholars in the 1930s had stressed the importance of studying the effect of ethnic lineage on migration and on shaping distribution of TIPs within and between indigenous communities and urban localities [19]. The importance of successfully computing TICD archive lies in the fact that quantitative study on migration of TIPs within/between indigenous communities and urban localities at individual level in the context of ethnic lineage becomes feasible.

Consequently, TICD is expected to help achieve the following goals: (1) to associate contemporary ethnic lineages with past ones in corresponding indigenous communities; (2) to construct contemporary ethnic lineage framework associated with indigenous communities and urban localities; (3) to link ethnic lineage of urban localties with that of the indigenous communities that serve as main origin of migration to urban localities.

Findings from archives of population dynamics suggest that TIPs are associated with much higher fertility and lower life expectancy, with infant and youth mortality rates being of much higher than the average level. TIPs infant and youth mortality rates vary by ethnic groups. Nevertheless, the author finds that based on population dynamics archives, TIPs with strong ethnic social network connection are associated with lower infant and youth mortality. This suggests that strengthening intra- and inter-ethnic relationship and connection might be an effective measure to overcome poverty trap and improve inequality, and thus promote social mobility.

One important contribution of the research is making “invisible” populations “visible” through the processes of computing archives of address geocodes, population dynamics, individual-level migration flows, and TICD. The paradox of “highly propensity of migration but low social mobility” observed among TIPs is likely embedded in complex inter- and intra-ethnic relations and the differential level of social network connection strength among different ethnic TIPs. Findings from the archives discussed in the paper shed lights on this viewpoint.

In the end, based on computational archival science and data science, a rich body of big quantitative data has been constructed by the author in the past decade. Researches of the author in [4], [8], and this paper have achieved the initial calls in [19] to a large extent.

ACKNOWLEDGEMENT

The research is mainly supported by a grant from the 2013-2021 bilateral joint research program of Academia Sinica and the Council of Indigenous Peoples on “Research on Contemporary Taiwan Indigenous Peoples”, and partly by the Ministry of Science and Technology (MOST 106-2420-H-001-008-MY2; 109-2420-H-001-003). The author appreciates the excellent research assistance from Debby Tien.

REFERENCES


