

Contemporary Chinese Village Gazetteer Data

数字村庄 數字村莊

East Asian Library, University of Pittsburgh

Yuanziyi Zhang*, yuz172@pitt.edu

Introduction

In July 2018, the East Asian Library (EAL) of the University of Pittsburgh Library System (ULS) initiated the Contemporary Chinese Village Data (CCVD) project to create an open-access online dataset of statistics selected from the library's collection of Chinese village gazetteers. This unique initiative has produced a dataset of significant value to the humanities and social sciences based on Chinese village gazetteers, which include quantitative and qualitative data critical to supporting Chinese studies in fields such as politics, economics, sociology, environmental science, history, and public health.

The first 500 villages' data were opened for access in October 2019. In March 2020, a dataset including statistics from 1,000 villages will be made available via the ULS digital collections website.

Data Entry

The categories of data we acquire are distributed in different chapters of a village gazetteer. To reduce mistakes and to ensure the accuracy of data, we set up a data dictionary containing standards and clarifications and follow the 3-step procedure.

1. **Pagination**
Browse the gazetteer and assign page numbers to different categories of data.
2. **Data Input**
Scrutinize the content on each labeled page, retrieve the data, and enter the data to the web forms strictly following the data dictionary
3. **Checking**
Double Check the data entered and make necessary adjustments.

CCVGD Website

<http://www.chinesevillagedata.library.pitt.edu/>

The dataset containing information on 800 villages is available for download in .csv format. An interactive map with basic administrative information on each collected village is also included in the website. The first 800 books cover 31 provinces/province-level municipalities/special administrative regions.

There are currently 18 datasets included in CCVG data within each there are multiple subcategories. The CCVG data is bilingual (Chinese/English).

Economy – Range

Economy – Yearly

Education – Range

Education – Yearly

Ethnic Groups – Range

Ethnic Groups – Yearly

Family Planning – Range

Family Planning – Yearly

Gazetteer Information

Last Names

Military, Politics and Management – Range

Military, Politics and Management – Yearly

Natural Disasters

Natural Environment

Population and Migration – Range

Population and Migration – Yearly

Village Information

Year of First Availability



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	村志代	村志书名	Category	Division1	Division1	Division1	Division1	19	19	19	19	19	19	19	19	19	19
35	6	方等村志	人口 Population	女性人口 Female	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
36	7	滹津洼村志	户数 Number of House	n/a	n/a	n/a	n/a	157	160	165	168	170	171	173	175	178	
37	7	滹津洼村志	人口 Population	女性人口 Female	n/a	n/a	n/a	325	338	340	348	350	356	362	366	370	
38	7	滹津洼村志	人口 Population	总人口 Total Popu	n/a	n/a	n/a	660	678	685	698	705	711	720	730	744	
39	7	滹津洼村志	死亡人数 Number of D	n/a	n/a	n/a	n/a	4	5	3	6	6	7	5	6	8	
40	7	滹津洼村志	人口 Population	男性人口 Male Pop	n/a	n/a	n/a	335	340	345	350	355	355	358	364	370	
41	7	滹津洼村志	出生人数 Number of B	n/a	n/a	n/a	n/a	8	10	14	11	10	10	9	13	10	
42	7	滹津洼村志	迁入 Migration In	n/a	人数 num	n/a	n/a	3	15	14	18	14	12	15	18	20	
43	7	滹津洼村志	迁出 Migration Out	n/a	人数 num	n/a	n/a	20	2	5	10	11	9	10	15	7	

Application

Research

As the size of the current datasets is still relatively small and the database is under design, there are few ongoing research that we know of based on the existing datasets. Below are quotes of feedback we received from faculty and scholars.

"In my research, I have used village level survey data from the Chinese government. CCVG provides the information that field surveys usually do not cover. I want to thank Pitt East Asian library for creating this very valuable dataset for the researchers."

Dr. Yifan Zhang

*Associate Professor, Department of Economics
The Chinese University of Hong Kong*

"I really appreciate that the Pitt Library is creating the datasets and making it publicly accessible. These data at the village level are of great value to people like me who want to understand the vast regional variation within China. It is excellent that the raw datasets are provided."

Yu Xiao

Independent scholar – Political Science

Teaching

Course Project Example:

While more and more arable land in China is requisitioned and factories are constructed instead, it is largely concerned whether this change would affect the gross output value in China. The project will investigate the relationship between arable land area and gross output value on the village level, which is the most basic administrative division of China. By analyzing the data, an understanding of the change in China's economic structure will be developed.

Yuanziyi Zhang, Graduate Student

LIS2030 Data and Information Systems

Database Construction

The East Asian Library is currently collaborating with the School of Computing and Information at the University of Pittsburgh to construct a database with tools of filtering, cross-searching, and visualizations.

Dr. Daqing He and a graduate student Qi Xin are developing the database with MySQL. The size of the current version is 7M. Below are examples of features that are already realized.

1. Refer to all the natural disasters in a specific village between 1998 and 2000.
2. Refer to the data of the population of a specific ethnic group which is available in 2005.
3. Refer to the villages that are in the same range of altitude.

Examples of features that will be realized in the near future:

1. Refer to all 18 categories of data of a specific village.
2. Refer to multiple categories of data of multiple villages.

Future Work

1. The project will continue extracting data from more gazetteers as time and resources allow with the goal of eventually reaching 2,500-3,000 village gazetteers.
2. Improve the database according to feedback
3. User Interface Design
 - JavaScript
 - Integrate Tableau to achieve visualization.
 - Provides drop-down filters for different categories on the same page for users to choose.

