華東師範大學
East China Normal University

东北三省民族三谈

——从少民政策优待生数据看民族

姓　　名：蒋尚辰

学　　号：10191511437

班　　级：

指导教师姓名：蒲鹏

指导教师职称：

2020 年 6 月

目　录

摘　要

2020 年正值第七次全国人口普查之际，本文尝试通过分析政府公布的享受照顾政策录取的少数民族考生名单，初步了解近 10 年来东北三省少数民族发展状况，并简单推广三省少数民族文化。文章的核心涵盖三个方面：一、东北三省少数民族状况概述；二、东北三省主要少数民族地理分布；三、满族、蒙古族、朝鲜族姓氏研究。

本文基于 Python 爬虫技术获取数据，使用 Excel 与 Python 作为处理、分析的工具，并结合 Excel 与 ECharts 展示可视化结果。

本文最后总结了东北三省少数民族发展的状况，并对我国今后少数民族事业做了期望和展望。

关键词：少数民族, 东北地区, Python, 可视化

一、 选题

(一)背景

2010 年，中国进行了第六次全国人口普查。今年，中国将迎来第七次全国人口普查。届时，关于各民族的统计数据将向人们详细展开。这 10 年来的民族发展对比将备受瞩目。

(二)目的

本文作者怀揣对数据的敬畏之心和对家乡的热爱之情，尝试通过享受照顾政策录取的少数民族考生数据，尽量还原且向各族同胞介绍东北地区 10 年来少数民族的发展状况，并简单推广东北地区特色民族文化。

二、 数据获取

(一)数据来源

本文数据来源于吉林省教育考试院[1]、黑龙江省招生考试信息港[2]、辽宁招生考试之窗[3]官方网站发布的享受照顾政策录取的少数民族考生公示名单。

(二)获取手段

本文作者使用 Selenium + XPath 从官方网站中提取出含有关键词“照顾”、“加分”的网页集，再从中获取数据（见**数据获取代码**）。

(三)获取结果

如表 1 所示，共收集原始数据集 13 个，数据 131951 条。

表 1 数据获取结果表

数据集	数据条数	数据集	数据条数
吉林 2017.xlsx	14192	辽宁 2018(2).htm	14003
辽宁 2015(2).htm	15049	辽宁 2018.htm	3221
辽宁 2015.htm	4108	辽宁 2019(2).htm	16003
辽宁 2016(2).mht	14365	辽宁 2019.htm	3299
辽宁 2016.mht	3960	黑龙江 2013.xls	13640
辽宁 2017(2).htm	14089	黑龙江 2014.xls	12455
辽宁 2017.htm	3567	**共计**	**131951**

三、 数据预处理

(一)脏数据处理

1. 吉林省数据

如表 2 所示，“照顾加分项目”属性中，含有非加粗项的记录为脏数据，经由 Excel 数据筛选清除。

表 2 吉林省数据示意表

姓名	照顾加分项目	毕业学校名称	民族	照顾加分分值
	散居地区不使用本民族语言文字答卷的少数民族考生	长春 XX 中学		
	少数民族聚居地区不使用本民族语言文字答卷的本民族考生	东北师大附中		
	高考用本民族语言文字答卷的少数民族考生	吉林 XX 中学		
	自主就业的退役士兵	永吉 XX 中学		
	立功军人子女、残疾军人子女、因公牺牲军人子女	舒兰 XX 中学		
	归侨、归侨子女、华侨子女	延边 XX 中学		
	・・・	・・・		

2. 黑龙江省数据

如表 3 所示，“照顾类别名称”属性中，含有非加粗项的记录为脏数据，经由 Excel 数据筛选清除。

表 3 黑龙江省数据示意表

报名号	姓名	考生特征代码	照顾类别名称	获奖项目和民族	考生所在学校班级
		01XXXX	**八少民族**		哈市 XX 中学
		02XXXX	**其他少数民族**		齐齐哈尔市 XX 中学
		03XXXX	**使用本民族文字答卷的少数民族考生**		鸡西市 XX 中学
		04XXXX	自谋职业的城镇退役士兵		鹤岗市 XX 中学
		05XXXX	侨眷考生		双鸭山市 XX 中学
		06XXXX	获国家二级运动员（含）以上称号者		大庆市 XX 中学
		・・・	・・・		・・・

3. 辽宁省数据

每一年度辽宁省涉及少数民族优待政策的名单共有两份，分别是“具备加分资格的少数民族考生名单”和“具备加分资格的‘双语生’考生名单”。两份表格的结构均如表 4 所示。脏数据的处理分为两步：第一步，合并两表，使用 Excel 依照考生号删除重复数据；第二步，“毕业中学”属性中，含有“往届生”字样的记录为脏数据，经由 Excel 数据筛选清除。

表 4 辽宁省数据示意表

序号	考生号	姓名	性别	民族	毕业中学
	XXXX03XXXX				鞍山市 XX 县高级中学
	XXXX04XXXX				抚顺市 XX 县往届生
	・・・				・・・

(二)数据降维、归一化

如表 2，吉林数据表中“毕业学校名称”按照地级行政区排列，因此可使用 Excel 快速转换成所在地级行政区（作者是吉林人，对吉林各地均有了解）。最后，属性更名为“地区”。

如表 3，黑龙江数据表中“考生特征代码”前两位决定考生所在地级行政区，参照“考生所在学校和班级”可快速转换成所在地级行政区。最后，属性更名为“地区”。

如表 4，辽宁数据表中“考生号”第 5-6 位决定考生所在地级行政区，参照“毕业中学”可快速转换成所在地级行政区。最后，属性更名为“地区”。

特别指出，黑龙江数据表中“考生特征代码”前两位为“40”和“41”的考生分别来自“农垦系统”和“森工系统”，并不属于某一地级市。由于历史特殊原因，农垦总局和森林工业总局管辖地遍布黑龙江全省，其数据在本表中只占 3.7%，故将其剔除。

(三)结果

经过前两步的处理，得到的表结构为：“姓名、民族、地区”。其他无关属性均被删除。最终共有数据集 3 个，数据 92933 条。

四、 数据分析与可视化

(一)黑、吉、辽少数民族状况概述

1.数据分析

使用 Excel 数据透视表对 92933 条数据进行分类汇总，其结果如下表 5 所示。

表 5 东三省少数民族大致占比表

吉林省			黑龙江省			辽宁省		
民族	**占比**	**人数**	**民族**	**占比**	**人数**	**民族**	**占比**	**人数**
满族	55.041%	7354	满族	54.072%	9402	满族	70.833%	44047
朝鲜族	27.094%	3620	朝鲜族	14.901%	2591	蒙古族	23.831%	14819
蒙古族	11.287%	1508	蒙古族	13.797%	2399	朝鲜族	4.590%	2854
回族	5.763%	770	回族	9.161%	1593	回族	0.357%	222
锡伯族	0.449%	60	达斡尔族	4.279%	744	锡伯族	0.326%	203
壮族	0.075%	10	锡伯族	1.098%	191	壮族	0.021%	13
苗族	0.060%	8	赫哲族	0.650%	113	鄂伦春族	0.005%	3
土家族	0.052%	7	鄂伦春族	0.477%	83	白族	0.005%	3
达斡尔族	0.045%	6	鄂温克族	0.397%	69	畲族	0.003%	2
黎族	0.045%	6	壮族	0.219%	38	达斡尔族	0.003%	2

彝族	0.037%	5	土家族	0.207%	36	土家族	0.003%	2
高山族	0.007%	1	柯尔克孜族	0.190%	33	彝族	0.003%	2
侗族	0.007%	1	苗族	0.184%	32	傣族	0.003%	2
畲族	0.007%	1	侗族	0.098%	17	苗族	0.003%	2
瑶族	0.007%	1	黎族	0.052%	9	傈僳族	0.003%	2
白族	0.007%	1	彝族	0.046%	8	黎族	0.003%	2
傣族	0.007%	1	布依族	0.040%	7	布依族	0.002%	1
佤族	0.007%	1	瑶族	0.035%	6	侗族	0.002%	1
总计	**100.000%**	**13361**	畲族	0.023%	4	高山族	0.002%	1
			俄罗斯族	0.017%	3	瑶族	0.002%	1
			维吾尔族	0.012%	2	**总计**	**100.000%**	**62184**
			白族	0.012%	2			
			高山族	0.006%	1			
			保安族	0.006%	1			
			水族	0.006%	1			
			藏族	0.006%	1			
			羌族	0.006%	1			
			京族	0.006%	1			
			总计	**100.000%**	**17388**			

2.可视化

表 5 在一定程度上反映了东北地区的民族省情。而从图 1、图 2、图 3 中可以看出，表 5 中的数据与第六次人口普查[4]的数据具有较高程度上的相似性。

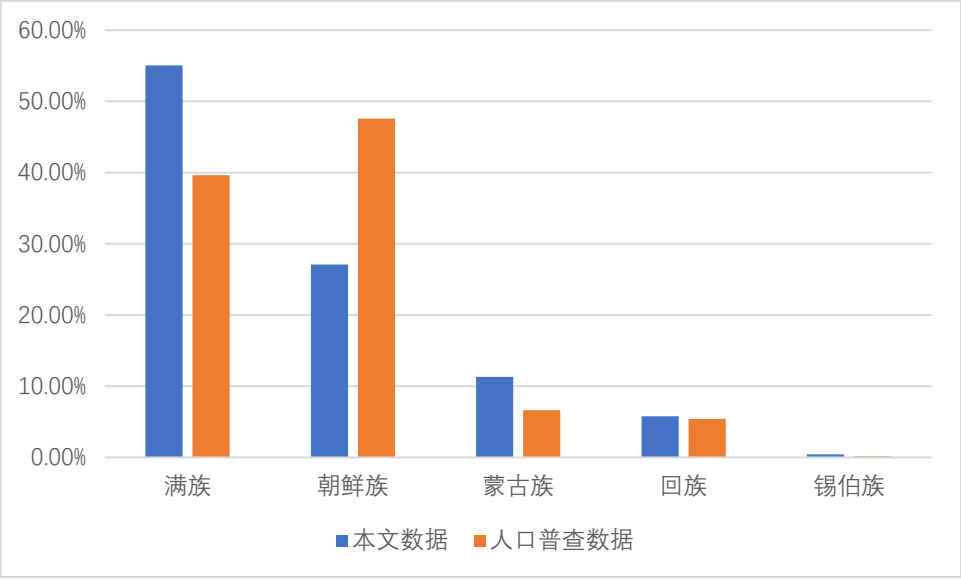

图 1 本文数据与人口普查数据对比图——吉林省

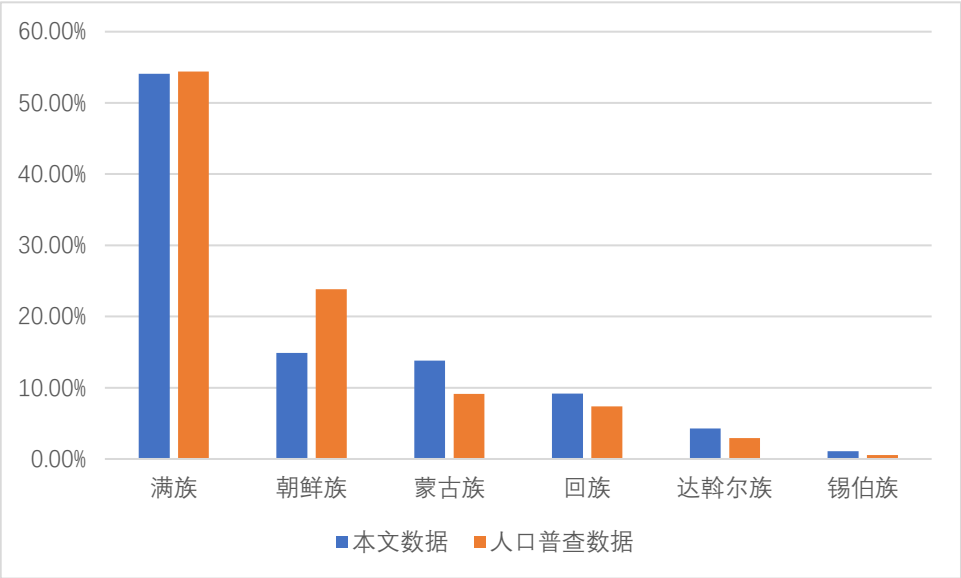

图 2 本文数据与人口普查数据对比图——黑龙江省

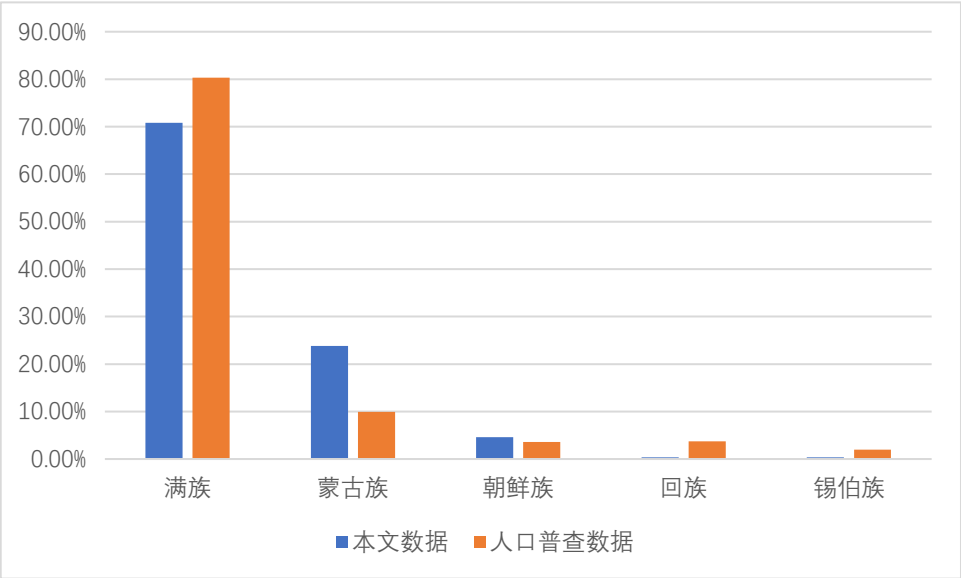

图 3 本文数据与人口普查数据对比图——辽宁省

3.综述

综合表 5 与图 1、2、3 可以看出，在人数上，东北三省的主体少数民族有满族（695 万人，占三省少数民族总人口 68%，本段括号中数据含义与此相同，本节括号中数据来自于第六次人口普查[4]）、朝鲜族（160 万人，占 15%）、蒙古族（92 万人，占 9%）、回族（46 万人，占 4%）。

东北三省中，黑龙江省是赫哲族（3613 人，占赫哲族总人口 67%，本段括号中数据含义与此相同）、鄂伦春族（3943 人，占 45%）、达斡尔族（40277 人，占 30%）、鄂温克族（2648 人，占 8%）的主要聚居地，辽宁省是锡伯族（132431 人，占 69%）、满族（533 万人，占 51%）的主要聚居地，吉林省是朝鲜族（104 万人，

占 56%）的主要聚居地。而除此之外，表 5 中其他民族属于散居在东北三省内的少数民族。

(二)黑、吉、辽主要少数民族地理分布

1.数据分析

根据上节概述，作者选取各地聚居民族数据绘制地理分布图。

使用 Excel 数据透视表得到某一民族在某一省份的地理分布，示例如表 6。

表 6 满族在辽宁省的分布表

满族	100.00%
鞍山市	17.40%
本溪市	14.38%
朝阳市	0.26%
大连市	0.00%
丹东市	33.63%
抚顺市	17.28%
阜新市	0.56%
葫芦岛市	0.00%
锦州市	16.49%
沈阳市	0.00%
铁岭市	0.00%
营口市	0.00%

依此原理得到各主要民族在三省的分布数据，并使用 Excel 中的 CONCAT 函数（代码为：=CONCAT("{name:'",A2,"', value:",B2,"},")）产生符合 JSON 标准的字符串，最终写入到 HTML 文件中（见**可视化代码**）。

2.可视化

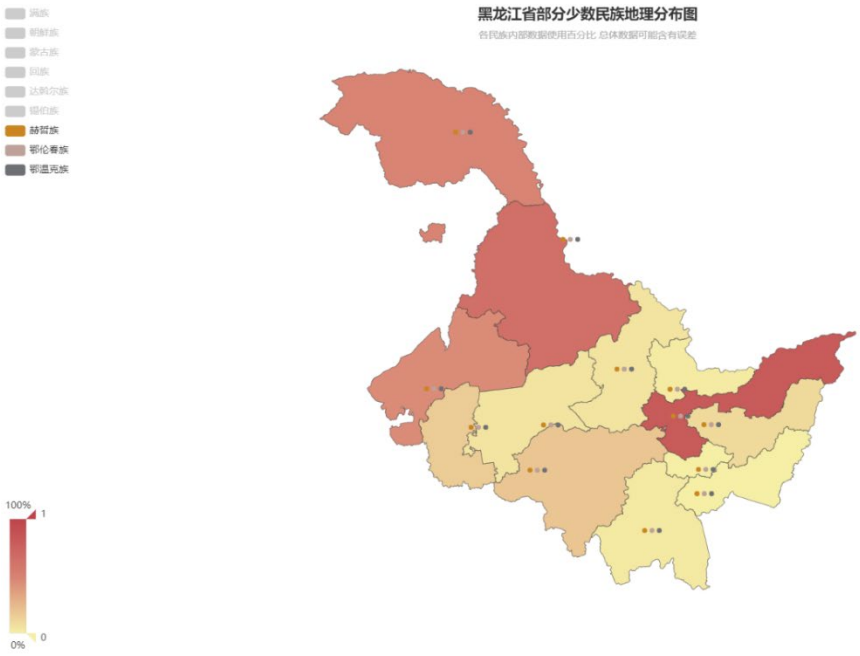

图 4 黑龙江省部分少数民族地理分布图(以赫哲族、鄂伦春族、鄂温克族为例)

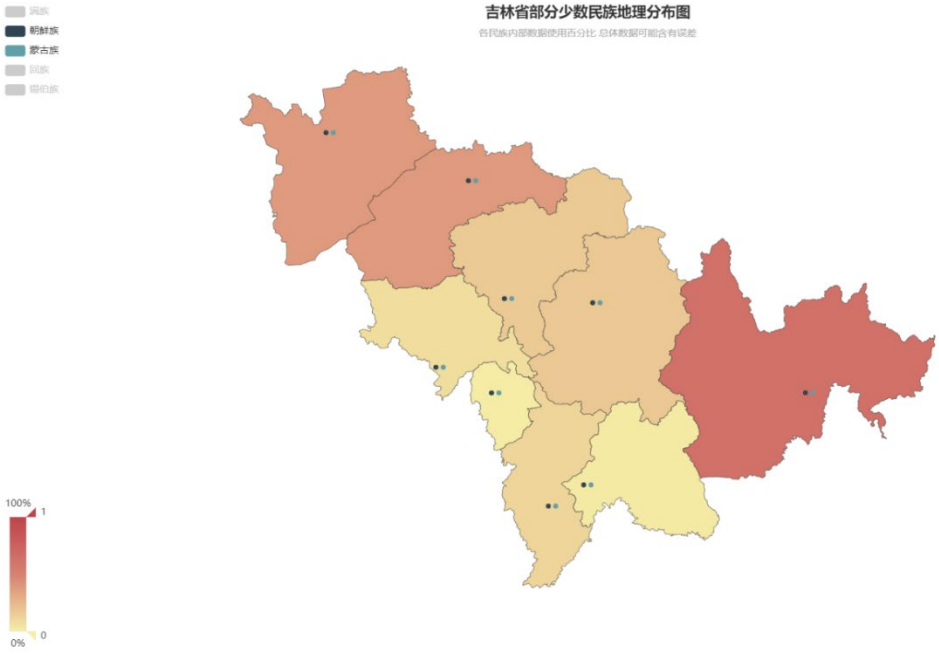

图 5 吉林省部分少数民族地理分布图(以朝鲜族、蒙古族为例)

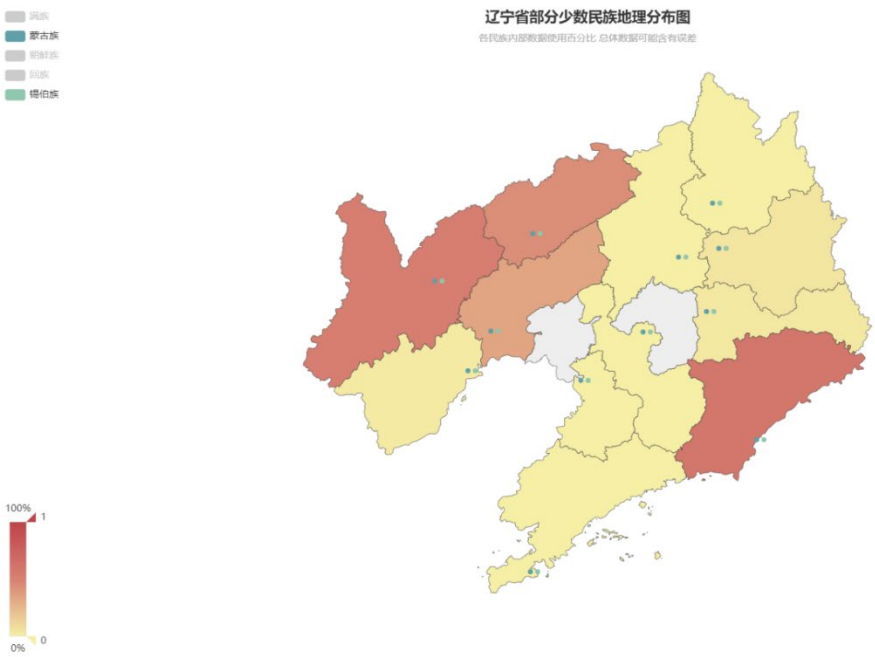

图 6 辽宁省部分少数民族地理分布图(以蒙古族、锡伯族为例)

(三)满族、蒙古族、朝鲜族姓氏研究

1.概述

由于满族、蒙古族、朝鲜族历史渊源殊异，其姓氏也各具特色。首先，关于朝鲜族姓氏，如朴尚春在《朝鲜族姓氏源流》中所说：“朝鲜族姓名全部使用汉字，其姓名结构同汉民族完全一致。”[5]因此，朝鲜族姓氏处理起来较为轻松。其次，关于满族姓氏，如今使用满族姓氏（如呼延氏、马佳氏等）者寥寥无几，大多数已改汉姓（如关姓、石姓等）或冠以汉字姓（如温特赫氏简称温、文扎氏简称文等）[6]。因此，满族姓氏也不难处理。最后，关于蒙古族姓氏，根据小林高四郎，传统蒙古族没有氏和姓[7]。而广为人知的一些蒙古族姓名只包含名而无姓氏，如乌兰其其格（红花）、斯琴高娃（聪明美丽）、义拉拉塔（胜利）、斯琴巴特尔（聪明的英雄）、查干巴拉（白虎）等。此外，还有汉姓加蒙名（如张舍楞）、汉姓加汉名或蒙汉两种姓名齐备等情况[8]。而从获取的数据中看，使用汉姓者占据了绝大多数。因此，简便起见，将无姓氏的蒙古族姓名的首字看作姓氏。

2.数据分析

第一步，使用 Python 截取包括复姓的所有姓氏（见**姓名分离代码**）。

第二步，使用 Excel 数据透视表将每一姓氏的民族人数汇总，其结果如表 7 前两列所示。由于姓氏人数在 100 以上的数据占总数据的 91.7%，故剔除人数在 100 以下的数据。使用组合 Excel 函数将某一姓氏中某民族人数替换为加权值，即考虑该人数

在该民族总人口中的比例，代码为：

=IF(MOD(ROW(B4),4)=1,B4/9065,(IF(MOD(ROW(B4),4)=2,B4/60803,IF(MOD(ROW(B4),4)=3,B4/18726,SUM(C5:C7)))))

统计每一姓氏中各民族占比，代码为：

=IF(MOD(ROW(C5),4)=1,C5/C4,(IF(MOD(ROW(C5),4)=2,C5/C3,IF(MOD(ROW(C5),4)=3,C5/C2,""))))

结果如表 7 所示。

表 7 姓氏民族汇总数据示意表

姓氏/民族	民族人数	加权人数	占比
王	**6730**	0.174485537	
朝鲜族	73	0.008052951	0.046153
满族	5116	0.084140585	0.482221
蒙古族	1541	0.082292	0.471626
李	**6600**	0.270017578	
朝鲜族	1245	0.137341423	0.508639
满族	4148	0.068220318	0.252651
蒙古族	1207	0.064455837	0.23871

第四步，使用组合 Excel 函数（=OFFSET(D5,(COLUMN(A1)-1)+(ROW(A1)-1)*4,,)）将表 7 中第 4 列的数据转置，并用=CONCAT("[",B2,",",C2,"]")函数进行组装，形成表 8。

表 8 姓氏民族数据示意表

	朝鲜族	满族	蒙古族	
王	0.046153	0.482221	0.471626	[0.0461525411868506,0.482220972886753]
李	0.508639	0.252651	0.23871	[0.508638823091642,0.252651396245167]
张	0.177649	0.420247	0.402104	[0.177648623562025,0.420246980146152]
刘	0.095926	0.487498	0.416576	[0.0959260449516025,0.487498118908977]
赵	0.2487	0.457364	0.293936	[0.248700022666215,0.457363523407402]

第五步，使用 Python 对表 8 中的数据进行 K-Means 聚类（见**聚类代码**）。

3.可视化

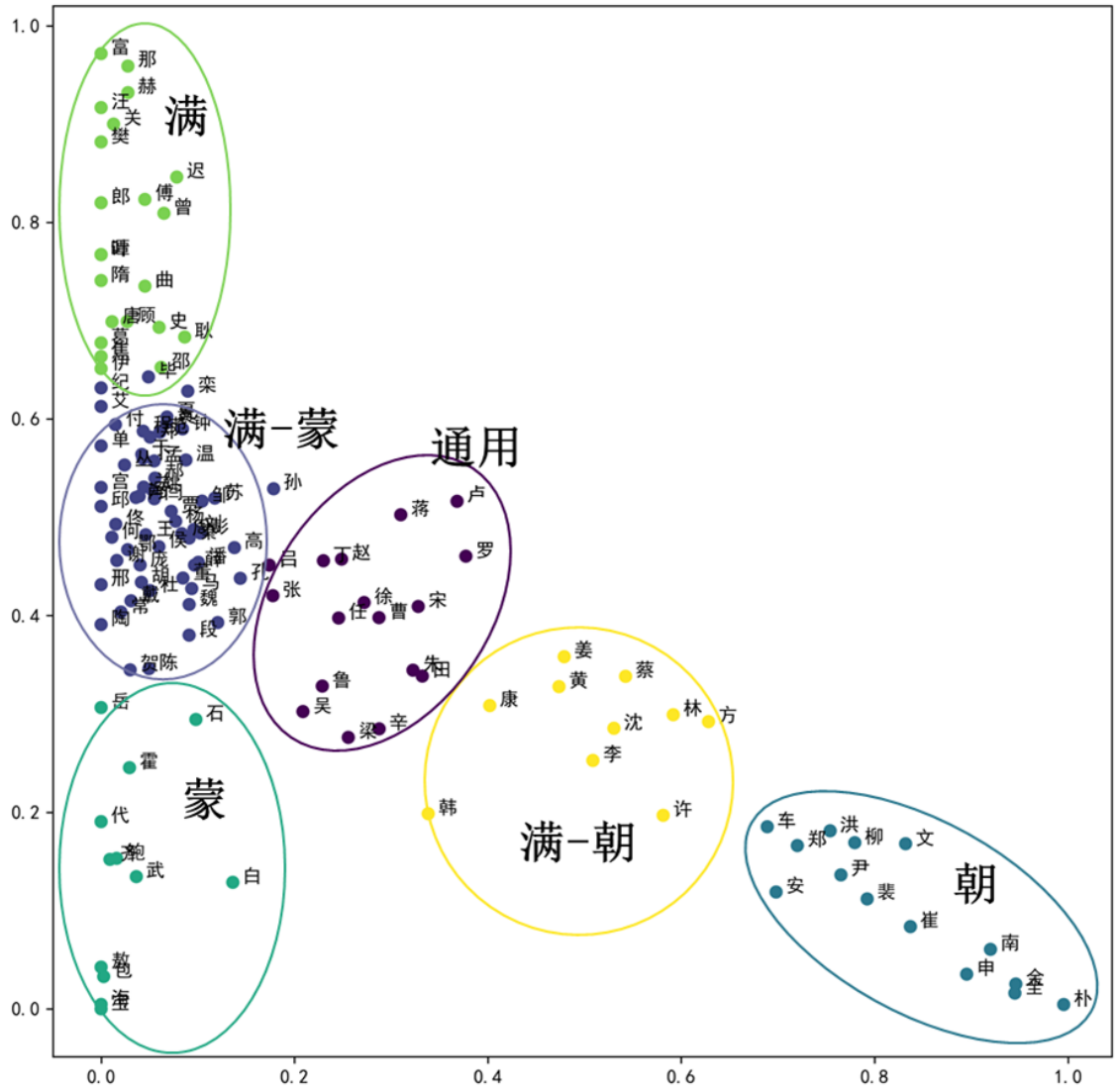

图 7 满族、蒙古族、朝鲜族姓氏关系图

如图 7 所示，姓氏数据共分为 6 个聚类。制图原理是对于同一姓氏，满族占比+蒙古族占比+朝鲜族占比=1，因此只需要使用两个变量就能在二维图片上呈现出三个民族各姓氏的关系。例如，朝鲜族姓氏“朴”的坐标靠近(1.0, 0)，代表该姓在满族和蒙古族中的比例接近 0；通用姓氏“辛”的坐标靠近(0.3, 0.3)，代表该姓在三个民族中的占比十分接近。

五、 结论与展望

总结起来，本文所做的三项工作依次是：一、利用人口普查数据较为成功地验证了少民政策优待生数据的可靠性；二、利用本文数据，简单地呈现了东北三省各自的民族分布情况；三、根据本文数据，摘选了三大民族的姓氏作为研究对象并初步揭示了三种民族姓氏的异同。

了解家乡的方式有千万种，数据分析也许是最有趣的一种。希望每个热爱家乡的人都能在数据分析中发掘出家乡文化可爱的一面。同时希望我国的少数民族事业也能跟紧时代步伐，用更具时代特色的方式发出每种文化的声音。

参考文献

[1] 阳光公示-吉林省教育考试院[E]. http://www.jleea.com.cn/ptgxzs/zsjh/1.html. 2020.

[2] 黑龙江招生考试信息港-阳光公示[E]. https://www.lzk.hl.cn/yggs/index_1.htm. 2020.

[3] 辽宁招生考试之窗[E]. http://www.lnzsks.com/listinfo/NewsList_1102_1.html. 2020

[4] 中国 2010 年人口普查资料[E]. http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm. 2010

[5] 朴尚春.朝鲜族姓氏源流[J].寻根,2009(02):93-96.

[6] 刘庆华.满族姓氏述略[J].民族研究,1983(01):64-71.

[7] 小林高四郎,乌恩.蒙古族的姓氏和亲属称谓[J].蒙古学资料与情报,1987(01):16-22.

[8] 乌日娜.蒙古族姓名趣闻[J].民族论坛,1986(03):76-77.

附录

数据获取代码（XPath、Selenium）

```
import re
from lxml import etree
from selenium import webdriver

chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument('--headless')
chrome_options.add_argument('--disable-gpu')


def url_list(url, n):
    ls = []
    for i in range(n):
        ls.append(re.sub(r'{i}', f'{i+1}', url))
    return ls


def url_finder(filename, url, *words):
    driver = webdriver.Chrome(chrome_options=chrome_options)
    driver.get(url)
    source_code = driver.page_source
    driver.implicitly_wait(15)
    driver.quit()
    html = etree.HTML(source_code)
    if len(words) == 0:
        path_l = f'//a[contains(text(),"{words[0]}")]/@href'
        path_t = f'//a[contains(text(),"{words[0]}")]/text()'
    else:
        raw_xpath = f'contains(text(),"{words[0]}")'
        for i in range(1, len(words)):
            raw_xpath += f' or contains(text(),"{words[i]}")'
        path_l = f'//a[{raw_xpath}]/@href'
        path_t = f'//a[{raw_xpath}]/text()'
    links = html.xpath(path_l)
    texts = html.xpath(path_t)
    i = 0
    with open(filename, 'a', encoding='utf-8') as file:
        for text in texts:
            text = re.sub(r'\s', '', text)
            file.write(text + '\n')
            r = re.split('/', url)
            l = r[0] + '//' + r[2] + links[i][links[i].index('/'):] + '\n'
```

```
            if l.index('//') == l.rindex('//'):
                file.write(l)
            else:
                file.write(links[i][links[i].index('/'):] + '\n')
            i += 1


def urls_finder(filename, url, n, *words):
    for u in url_list(url, n):
        url_finder(filename, u, *words)
```

可视化代码（ECharts）

```
//篇幅所限，仅保留满族数据。
var option = {
    title: {
        text: '辽宁省五大少数民族大致分布(百分比)',
        subtext: '数据含有一定误差',
        left: 'center'
    },
    tooltip: {
        trigger: 'item'
    },
    legend: {
        orient: 'vertical',
        left: 'left',
        data: ['满族']
    },
    visualMap: {
        min: 0,
        max: 1,
        left: 'left',
        top: 'bottom',
        text: ['100%', '0%'],
        calculable: true
    },
    toolbox: {
        show: true,
        orient: 'vertical',
        left: 'right',
        top: 'center',
        feature: {
            mark: { show: true },
            dataView: { show: true, readOnly: false },
            restore: { show: true },
            saveAsImage: { show: true }
```

```
        }
    },
    series: [
        {
            name: '满族',
            type: 'map',
            mapType: '辽宁',
            roam: false,
            label: {
                normal: {
                    show: false
                },
                emphasis: {
                    show: true
                }
            },
            data: [
                { name: '鞍山市', value: 0.173973255840352 },
                { name: '本溪市', value: 0.143800939905101 },
                { name: '朝阳市', value: 0.00256544146025836 },
                { name: '大连市', value: 0 },
                { name: '丹东市', value: 0.336345267555112 },
                { name: '抚顺市', value: 0.172838104751742 },
                { name: '阜新市', value: 0.00558494335596068 },
                { name: '葫芦岛市', value: 0 },
                { name: '锦州市', value: 0.164892047131473 },
                { name: '沈阳市', value: 0 },
                { name: '铁岭市', value: 0 },
                { name: '营口市', value: 0 }
            ]
        }
    ]
};
```

姓名分离代码

```
# -*- coding:utf-8 -*-
import xlrd

compound_name = ['欧阳', '太史', '端木', '上官', '司马', '东方', '独孤', '南宫
', '万俟', '闻人', '夏侯', '诸葛', '尉迟', '公羊', '赫连', '澹台', '皇甫', '宗
政', '濮阳', '公冶', '太叔', '申屠', '公孙', '慕容', '仲孙', '钟离', '长孙', '
宇文', '司徒', '鲜于', '司空', '闾丘', '子车', '亓官', '司寇', '巫马', '公西
', '颛孙', '壤驷', '公良', '漆雕', '乐正', '宰父', '谷梁', '拓跋', '夹谷', '轩
辕', '令狐', '段干', '百里', '呼延', '东郭', '南门', '羊舌', '微生', '公户', '
公玉', '公仪', '梁丘', '公仲', '公上', '公门', '公山', '公坚', '左丘', '公伯
```

```
', '西门', '公祖', '第五', '公乘', '贯丘', '公皙', '南荣', '东里', '东宫', '仲
长', '子书', '子桑', '即墨', '达奚', '褚师']

excel = xlrd.open_workbook('姓名.xlsx')
table = excel.sheet_by_name('总')
column = table.col_values(0)
file = open('姓氏.txt', 'a', encoding='utf-8')
for name in column:
    slice_name = name[:2]
    if slice_name in compound_name:
        file.write(f'{slice_name}\n')
    else:
        file.write(f'{name[:1]}\n')
file.close()
```

聚类代码（K-Means）

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['figure.figsize'] = (10, 10)

data = np.array([[0.0461525411868506,0.482220972886753],[0.508638823091642,0
.252651396245167],[0.177648623562025,0.420246980146152],[0.0959260449516025,
0.487498118908977],[0.248700022666215,0.457363523407402],[0.946056145627727,
0.0252649741212833],[0.178451170494837,0.528951878989152],[0.077159070258422
5,0.495964945614024],[0.208760630453786,0.302380787867758],[0.04954846875043
27,0.346011015153693],[0.0420927739269298,0.563960881461979],[0.136195107489
286,0.128683326735712],[0.0128700959908963,0.899906584442477],[0.27202199179
1498,0.413257744411647],[0.137869216590401,0.469102814188105],[0.47896493551
6755,0.35813829727272],[0.0937384351219869,0.427450787679929],[0.33829034784
7095,0.198497924653846],[0.837039937128827,0.0836352805571638],[0.0833590485
008959,0.483377450622848],[0.32800086408916,0.409254843036542],[0.1205995892
54694,0.392895339100801],[0.00265647268273323,0.0328720090147701],[0.4735838
6252885,0.327728073532554],[0.0111318165147361,0.479630378896732],[0.9955849
49079468,0.00441505092053226],[0.720098419412282,0.166076347144358],[0.32258
3112289263,0.344432779918792],[0.0846623031548997,0.438168459916445],[0.0148
175839406688,0.594254496909065],[0.00914430605329798,0.152008691719254],[0.0
549697746911588,0.557282839911257],[0.0416709059725281,0.433848915203102],[0
.015166457724085,0.492927960143497],[0.287410332779834,0.397887682544323],[0
.0980589834474004,0.294337686029863],[0.591753347123238,0.29911915321147],[0
.581350607729106,0.196849229036876],[0.0112958472002026,0.698891416064705],[
0.173963995544573,0.451285544812609],[0.255652554160204,0.276074317158018],[
0.332226732022689,0.338462050549075],[0.368156608564241,0.516241508579958],[
0.245691782482253,0.397500141868656],[0.0505867759316689,0.425362212119149],
```

```
[0.0453684214538253,0.735009267065385],[0.765166505521554,0.136236048775492]
,[0.0394428706504639,0.521400805720973],[0.698118167771468,0.118825886215349
],[0.0550200297599999,0.518828890650638],[0.402012650917112,0.30844740395067
3],[0.0607441668755606,0.586390807346191],[0.542446934896475,0.3381934614208
35],[0.0725354848773462,0.506103303707631],[0.0910871466226018,0.41128012160
2044],[0.117608548494414,0.519445573467062],[0.036440512705868,0.13446290281
8079],[0.229947126264702,0.455753806756083],[0.100591785554147,0.45419582596
3118],[0.0677594972028327,0.596025700812082],[0.0681462152802122,0.601967294
537622],[0.0597700180425474,0.470054972864213],[0.377271343144225,0.46038920
1287052],[0,0.916661265711151],[0.0162701275573097,0.456028103641568],[0.053
0618113346808,0.527392091123663],[0.754001572269996,0.181109213659229],[0.10
4434274025332,0.516401718429155],[0.0359637970105709,0.520091879189843],[0.5
3042162803711,0.285373920213449],[0.0277219402377643,0.958858248363361],[0.0
277800406660486,0.931876148273758],[0,0.190432507412404],[0.0435464628158278
,0.587549233278059],[0.0269489980788855,0.699092547404024],[0,0.431477824254
369],[0.689003411330897,0.185236694697335],[0.0205100327030569,0.40362967175
672],[0.0242497276285982,0.553147681625021],[0.0506558067521824,0.5815174644
68153],[0,0.740754900400218],[0,0.390846111074933],[0.0911145562938538,0.478
839189042208],[0.628220731558155,0.292130613087027],[0.0649975453174585,0.80
9144770542253],[0.0599913978224372,0.693159986311209],[0.0160413563487707,0.
153060758503739],[0,0.677403904821483],[0.779503974189086,0.169039590475328]
,[0.945075406626192,0.0159885887695984],[0.0620750708011016,0.65245286310289
],[0.055704047213972,0.539812463524048],[0.0877566940266118,0.55822819714464
9],[0.8320559420537,0.1679440579463],[0.0274790834453843,0.467035501920869],
[0,0.767062548458707],[0,0.0424484507485871],[0.895240538856415,0.0352561355
079201],[0.0782652943332513,0.845960392612074],[0,0.631535515257034],[0.2875
87133242796,0.284817824713387],[0,0.511064471536335],[0,0.819859865000321],[
0,0.530351832541614],[0,0.00450867135527396],[0.143969538582691,0.4378683764
27852],[0,0.612686553335202],[0.0452626760041959,0.823271076646225],[0.03093
1869155861,0.41504145273768],[0.0966097132425616,0.451305049373217],[0,0.881
699827344747],[0.102509707310688,0.483960754092625],[0.792315959146597,0.111
796716148929],[0.091045654082582,0.380066903265315],[0.0300340979504124,0.34
4784805681917],[0.0864742161567134,0.683290376813844],[0,0.767191022593609],
[0.228763979432931,0.32826998968818],[0,0],[0,0.651217702122778],[0.08417630
15945701,0.589834945247349],[0,0.306461875386513],[0,0.971579268660351],[0,0
.66341986189824],[0.0896563573732184,0.628234451271262],[0.0293795066082221,
0.245287448556779],[0.0431624264052856,0.559845129297247],[0,0.5725957412110
55],[0.309939100172426,0.502514310655898],[0.040374927052719,0.4514564006704
82],[0.0489887035511497,0.642719415108188],[0.0439408683166256,0.53063568696
4576],[0.91985244429911,0.0605024984804211]])
label = ['王', '李', '张', '刘', '赵', '金', '孙', '杨', '吴', '陈', '于', '白
', '关', '徐', '高', '姜', '马', '韩', '崔', '周', '宋', '郭', '包', '黄', '何
', '朴', '郑', '朱', '董', '付', '齐', '孟', '胡', '佟', '曹', '石', '林', '许
', '唐', '吕', '梁', '田', '卢', '任', '杜', '曲', '尹', '冯', '安', '闫', '康
', '范', '蔡', '贾', '魏', '苏', '武', '丁', '潘', '袁', '夏', '侯', '罗', '汪
```

```
', '谢', '姚', '洪', '邹', '肖', '沈', '那', '赫', '代', '程', '顾', '邢', '车
', '常', '丛', '邓', '隋', '陶', '秦', '方', '曾', '史', '鲍', '葛', '柳', '全
', '邵', '郝', '温', '文', '鄂', '谭', '敖', '申', '迟', '纪', '辛', '邱', '郎
', '宫', '海', '孔', '艾', '傅', '戴', '薛', '樊', '彭', '裴', '段', '贺', '耿
', '叶', '鲁', '宝', '伊', '钟', '岳', '富', '焦', '栾', '霍', '卜', '单', '蒋
', '庞', '毕', '兰', '南']

y_pred = KMeans(n_clusters=4, random_state=9).fit_predict(data)

plt.scatter(data[:, 0], data[:, 1], c=y_pred)
for i in range(len(data[:, 0])):
    plt.annotate(label[i], xy = (data[:, 0][i], data[:, 1][i]), xytext = (da
ta[:, 0][i]+0.01, data[:, 1][i])) plt.savefig('聚类结果.png', dpi=300)
plt.show()
```

致谢

感谢华东师范大学数据库高级应用与数据分析高级应用这个优秀的集体，感谢蒲鹏老师和各位同学，在大家的相互交流与沟通中，我不断地丰富了自己的阅历和见识，成为了百尺竿头更进一步的激流勇进者。

值此论文完成之际，特向诸位表示感谢。