Final Project: 从高等教育统计数据看高等教育发展与改革

17307110020

黄永晟

1. 引言

行政数据是来源于行政系统的数据，它们来自不同的行政部门，如教育、医疗、税务、住房等部门。行政数据为社会科学研究提供了新的机遇与挑战。本次数据可视化的项目将焦点放在高等教育的改革与发展上，我们着重展示了不同类别高等教育毕业生人口统计特征在时空多维度的发展规律与趋势。作为数据小白，通过该可视化作品可以增进对我国高等教育发展的认识，也可以回答诸如哪个省份去年本科毕业人数最多，高等教育男女性别比例如何变化等问题。作为教育研究人员，通过该可视化作品可以增加对高等教育毕业生人口统计特征数据的基本认识，通过交互式分析对数据进行探索。在该可视化作品中，你可以研究地域平衡分析，“阴盛阳衰”性别分析，发展趋势分析等科研问题。

2. 数据介绍

我们的数据来源与中华人民共和国教育部政府门户的教育统计数据 [i]。我们选择了各地基本数据中高等学校(机构)研究生数、高等教育普通本专科学生数、高等教育成人本专科学生数、高等教育网络本科、专科生学生数等可以全面反映高等教育人口统计特征的数据。由于 2013 年以前的数据地理颗粒度没有具体到省级行政单位，因此我们下载的数据是 2013 年-2019 年。我们主要下载的字段有年份、地区、毕（结）业生数（合计）、毕（结）业生数（女）、毕（结）业生数（本科）、毕（结）业生数（专科）。我们提取了本专比例、硕博比例、性别比例等计算特征，并且利用全国人口统计数据，计算了相应字段的全国平均水平，用于可视化与分析。下面两式展示了全国平均水平及省市期望水平的计算方式。

$$\overline{\text{男女比例}} = \frac{\text{女生}_{\text{全国}}}{\text{总量}_{\text{全国}}}$$

$$E[\text{北京研究生数}] = \frac{\text{人口}_{\text{北京}}}{\text{人口}_{\text{全国}}} \cdot \text{研究生数}_{\text{全国}}$$

3. 模块简介

3.1 地图

地图是研究社会数据时展示空间分布的最好的信息可视方式。本项目中，我们利用中国地图展现社会数据在省级行政区上的分布情况。在图 1 中，我们展示

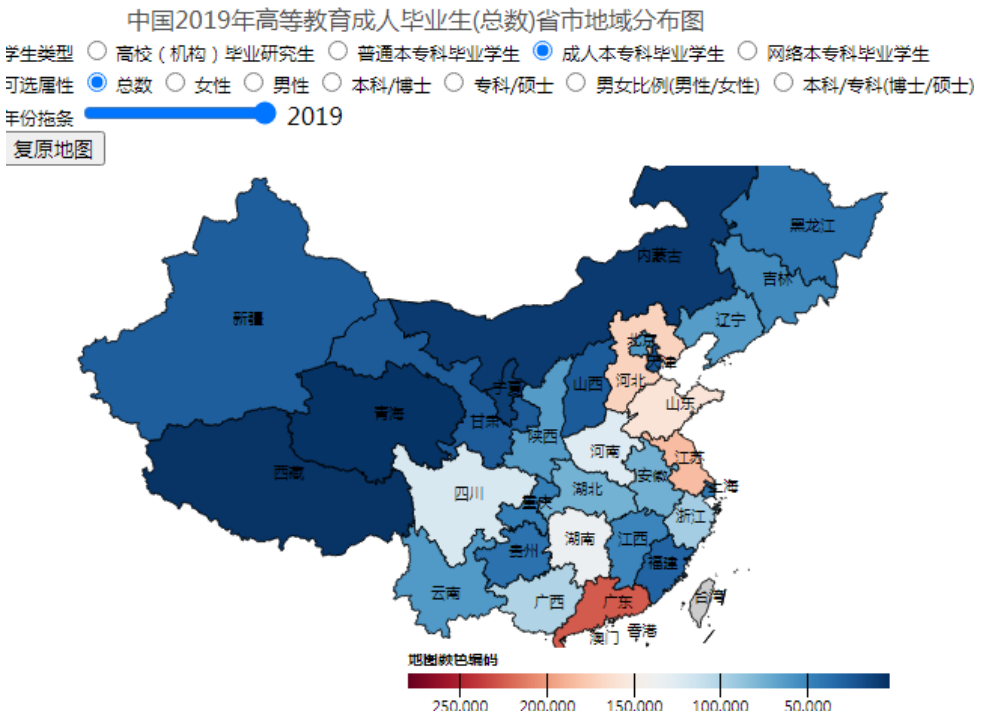

图 1 中国 2019 年高等教育成人本专科毕业生总数的省市分布地图

了中国 2019 年高等教育成人本专科毕业生总数的省市分布情况。从图 1 可见，广东毕业生总数最多，河北，山东，江苏次之，而西藏，青海，内蒙古等地区人数最少。利用 3.3 节的聚类颜色编码交互分析，我们利用 34 个省市高等教育成人本专科毕业生总数及时序发展的不同规律对地图颜色重新编码，如图 2。

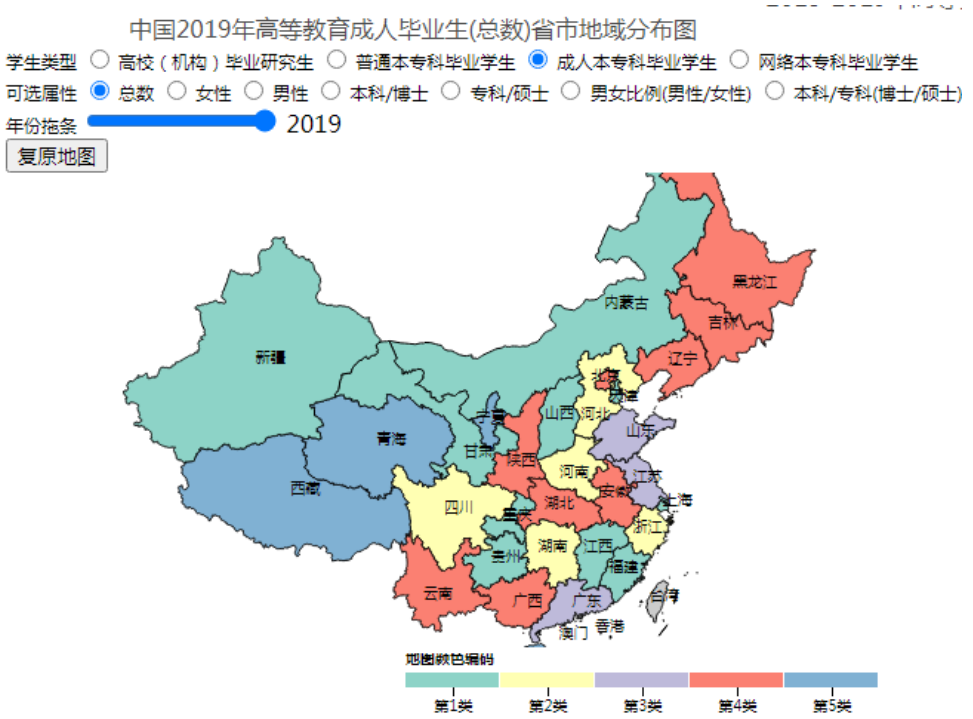

图 2 中国 2019 年高等教育成人本专科毕业生总数的省市分布地图（聚类编码）

从图 2 可见，山东，江苏与广东被聚为一类，而青海与西藏则被聚为一类，我们要利用 3.3 节的可视化视图来辨别类与类之间的区别。在任意的颜色编码下，我们都可以对通过悬浮鼠标了解某省市该属性的具体数值。地图支持缩放、平移、复原等基本操作。

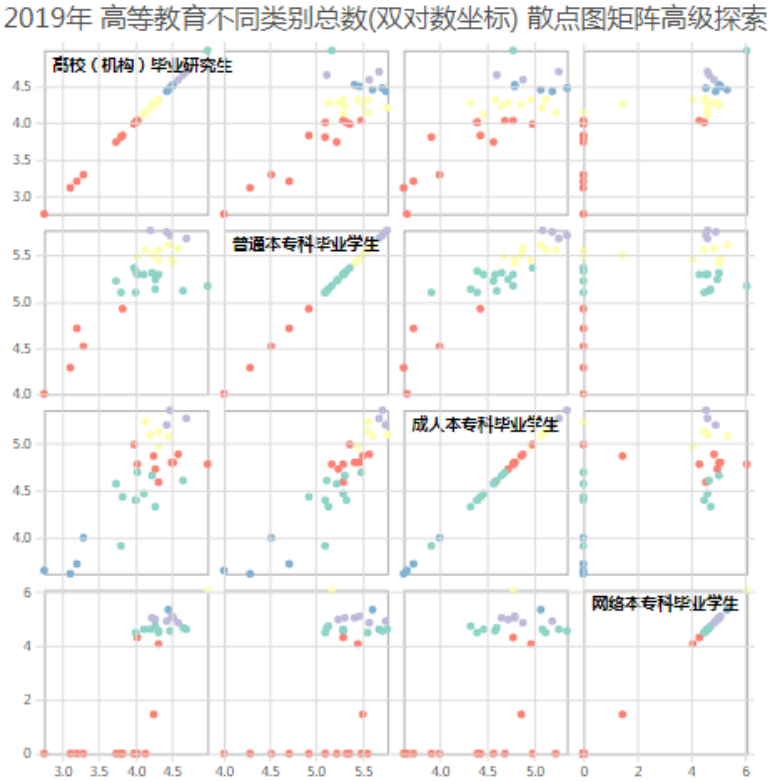

图 3 2019 年高等教育不同类别总数(双对数坐标)散点图矩阵

3.2 散点图矩阵

散点图矩阵是展示多个变量二元关系的绝佳方式。本研究中，我们想要了解不同高等教育类别（研究生，普通本专科生，成人本专科生，网络本专科生）数据之间的相互联系。与传统的散点图矩阵颜色编码方式不同，散点图矩阵的每一行颜色编码不同，如图 3。散点图矩阵的每个元素代表 2019 年一个省级行政区的数据，矩阵的第一行反映 2019 年高校机构毕业研究生总数与其他类别毕业生总数的相关关系，颜色编码采用高校机构毕业研究生总数的聚类编码；第二行的颜色编码则采用了普通本专科学生的总数的聚类编码，以此类推。我们研究一个有趣的问题，某个省市的性别比例情况在不同高等教育类别里是否具有某种联系？

从图 4 中，我们首先发现了一个离群点，安徽省研究生与网络本专科学生男生远多于女生。同时我们通过坐标轴的尺度与散点的分布发现在研究生，普通本专科生与成人本专科生三个类别中男女比例均小于 1。这表明“阴盛阳衰”的现象在某种程度上是现今高等教育中普遍存在的。

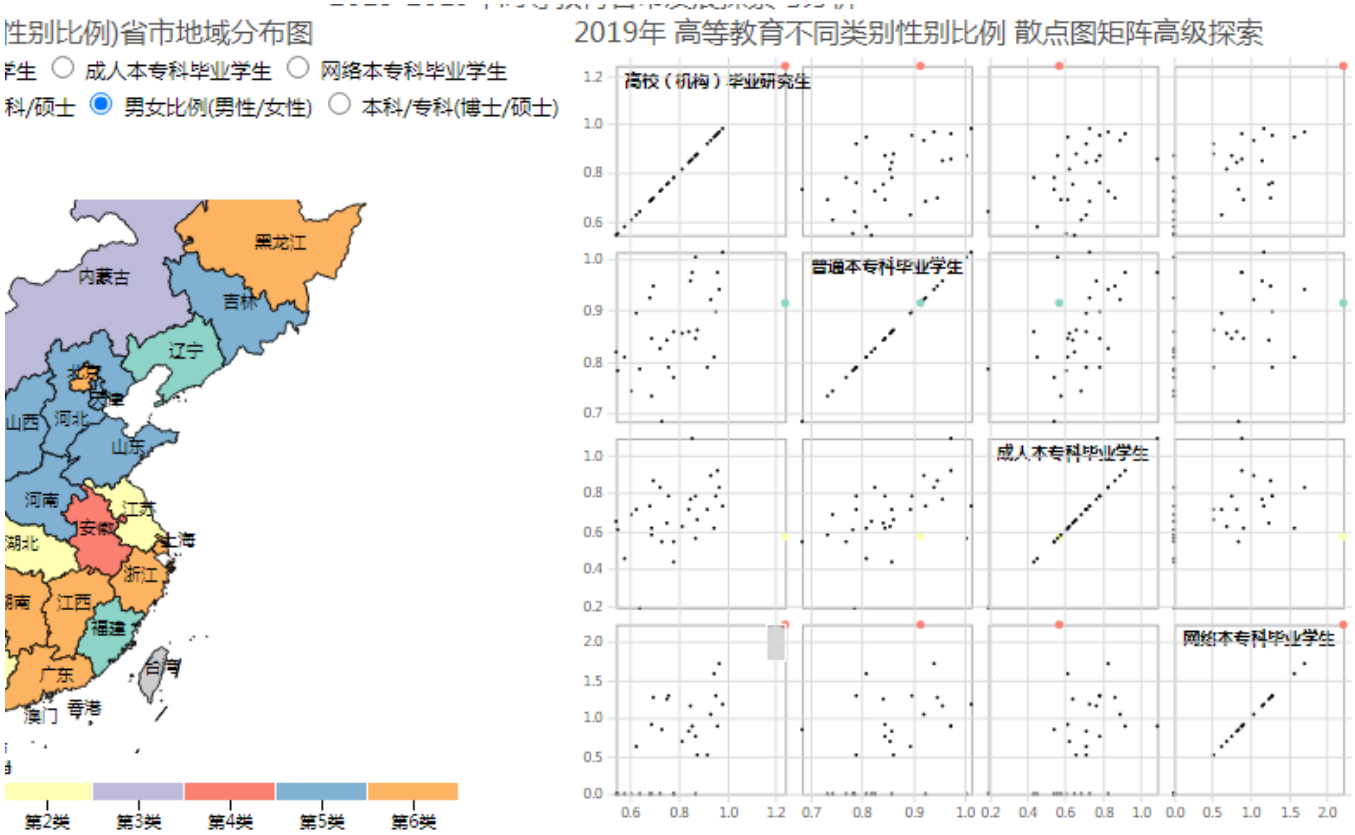

图 4 2019 年高等教育不同类别性别比例散点图矩阵

在图 5 中我们重新刷选相关数据。我们可以发现普通本专科生中男性占比偏高的省市在其他教育类别中男女占比也普遍偏高，这种联系体现在与研究生男女占比的关系上尤为显著。通过与其他视图的交互，我们发现西藏，吉林与天津市唯独的普通本专科生中男性偏高，但研究生中女性占比偏高的省市，其中的原因有待研究。

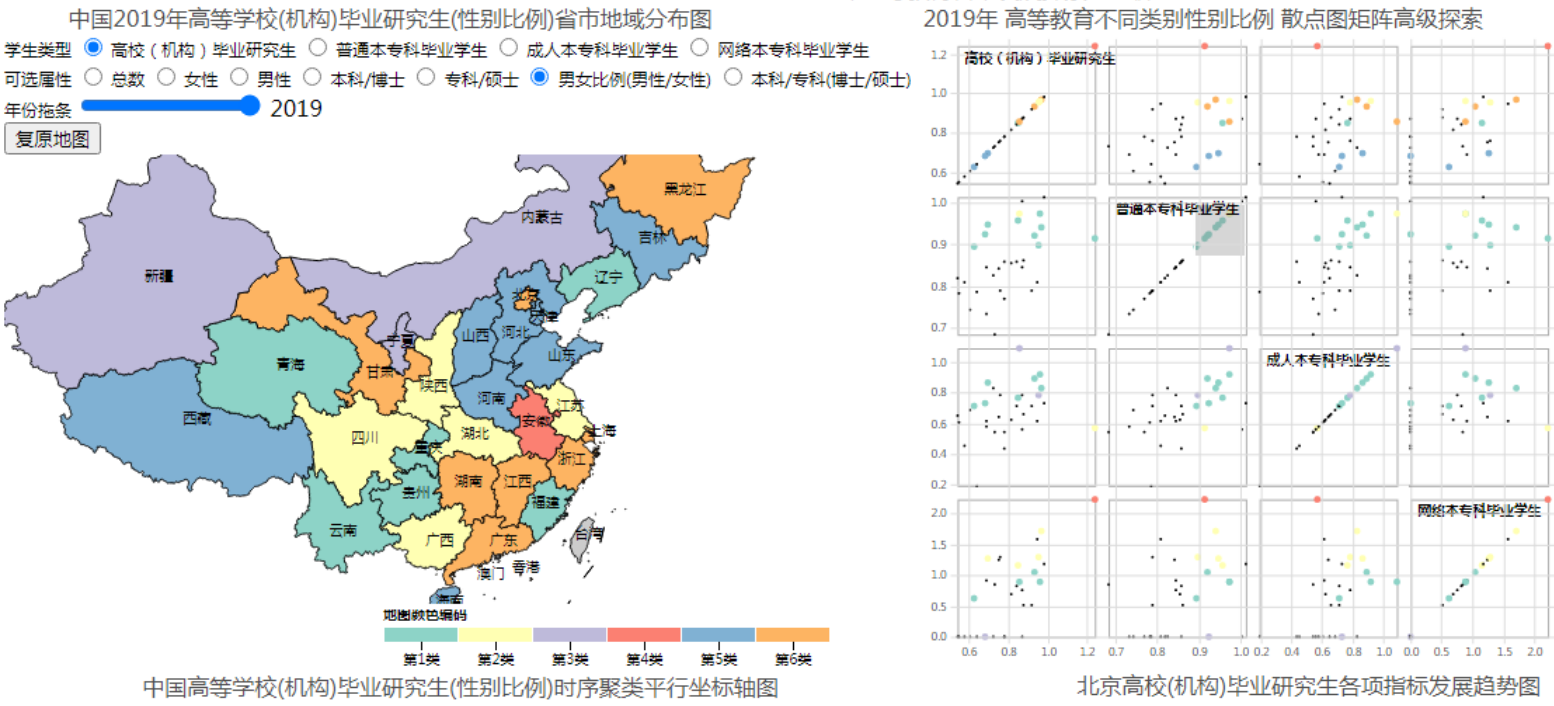

图 5 2019 年高等教育不同类别性别比例散点图矩阵 2

3.3 平行坐标轴图

平行坐标轴图是可视化高维多元数据的一种常用方法。本研究中我们利用平行坐标轴可视化不同年份某个人口统计特征的时序变化规律。我们利用 DTW 相似度对时间序列进行 K-Means 聚类，利用聚类结果来对平行坐标轴的颜色进行编码，我们在图 6 中展示了中国高等学校(机构)毕业研究生(总数)时序聚类平行坐标轴图的基本情况。从图 6 可以见到 34 个省级行政区被聚为 5 类，其中绿色折线表示的北京市一枝独秀，紫色折线表示的湖北，江苏和上海紧随其后。通过平行坐标轴+聚类算法的数据可视分析算法，我们对不同地域人口统计特征的变化趋势的异同有更清晰的认识。

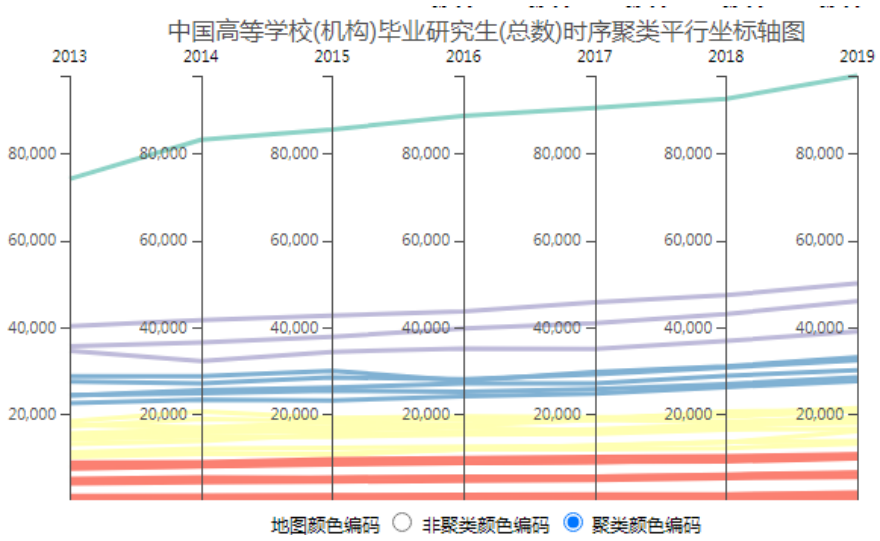

图 6 中国高等学校(机构)毕业研究生(总数)时序聚类平行坐标轴图

3.4 时间序列折线图

在详细视图中，我们利用时间序列折线图展示某一个省市不同人口统计特征的随年份变化的规律，图 7 进行了视图的简单展示。从图中，我们可以发现北京高校机构毕业生总数，男女比例及硕博比例均高于全国的均值。毕业生总数稳步

增加，女生占比逐年下降，硕博比例稳定在 20%上一个较高的水准。

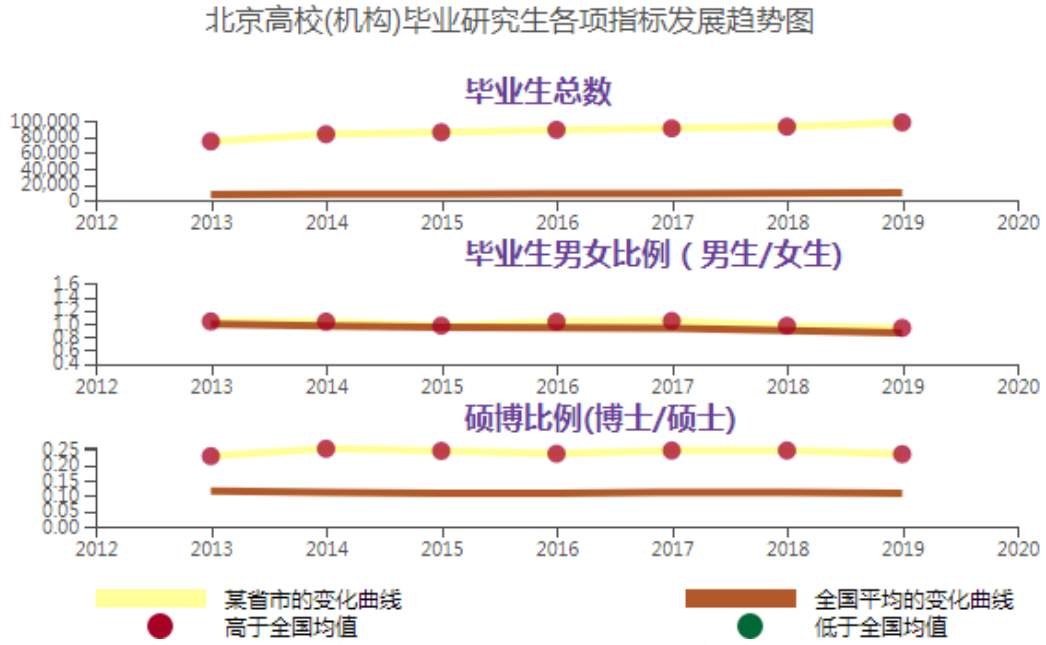

图 7 北京高校(机构)毕业研究生各项指标发展趋势图

4. 交互展示

4.1 单选框与时间拖条

利用单选框与时间拖条，我们可以选择不同高等教育的类型，不同的人口统计特征属性以及不同年份从各个角度筛选数据进行可视化。学生类型，可选属性与年份拖条与地图进行交互；可选属性与年份拖条与散点图矩阵进行交互；学生类型与可选属性与平行坐标轴图进行交互，见图 8。同时在地图的颜色编码中我们给出了聚类颜色编码（离散）与非聚类颜色编码（连续）两种方式，见图 9。

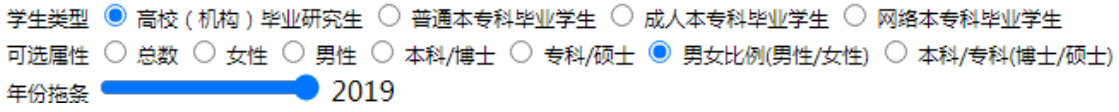

图 8 单选框与时间拖条

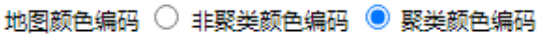

图 9 颜色编码单选框

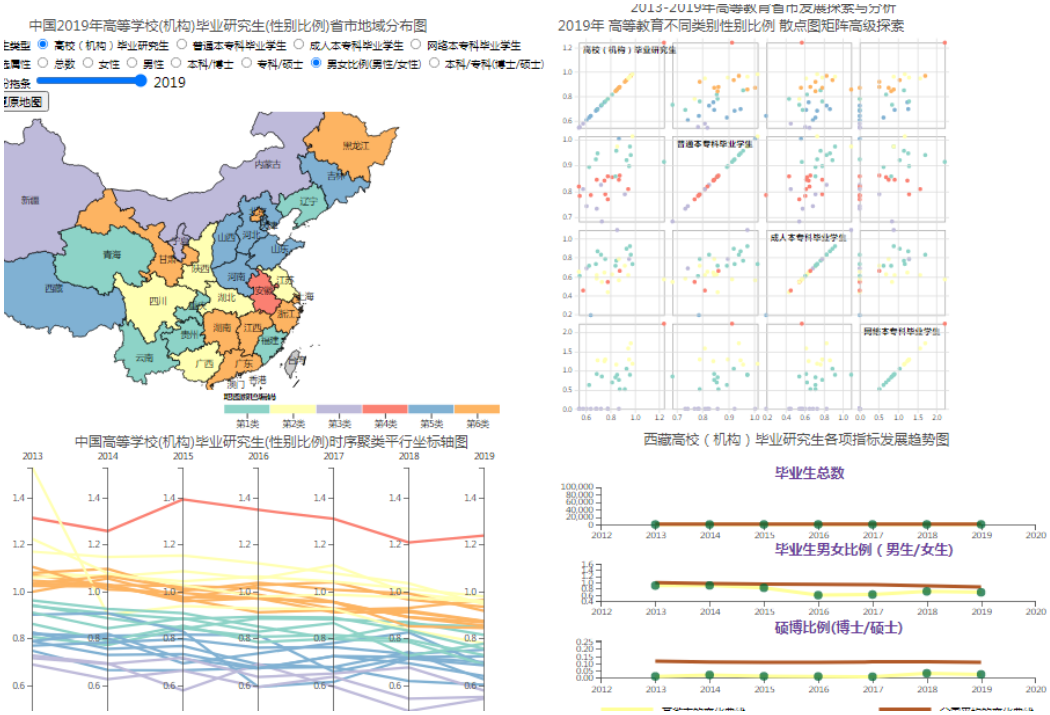

图 10 共用颜色编码

4.2 共用颜色编码

地图，平行坐标轴图与散点图矩阵共用一套颜色编码。从散点图矩阵我们可以刷选感兴趣的数据点，下方会显示刷选的省市，通过与左侧平行坐标轴图的交互我们可以观察到刷选省市人口统计特征的时序变化规律，如图 10。

4.3 散点图矩阵-平行坐标轴与地图的交互影响

刷选散点图矩阵感兴趣的数据点，左侧的平行坐标轴图会高亮在单选框选中的学生类型与人口统计特征下的刷选省市的时间序列，而地图也会高亮这部分省市所对应的地图区域，如图 11。通过三个视图的交互，我们可以同时了解某些省市的地理布局，不同教育类别人口统计特征的相关性及随时间变化的时序变化规律。

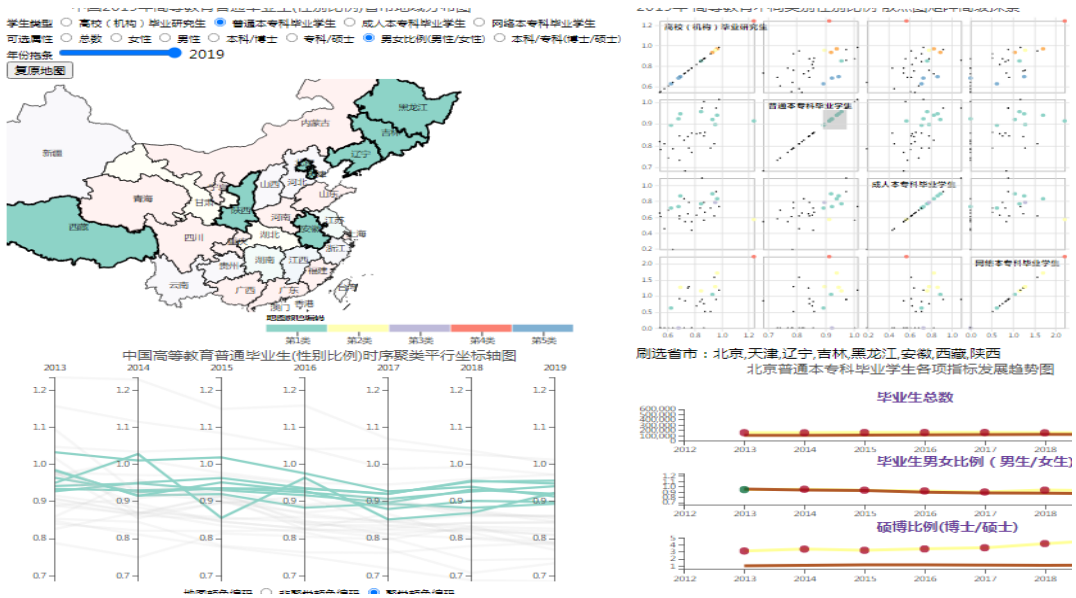

图 11 散点图矩阵-平行坐标轴图与地图的交互影响展示

4.4 平行坐标轴与地图的交互影响

取消散点图矩阵的刷选框，通过悬浮鼠标在平行坐标轴图某一聚类的时间序列上时，地图上会高亮出对应的省市区域。同时，我们可以更直观地比较聚类结果中不同类别的差异。

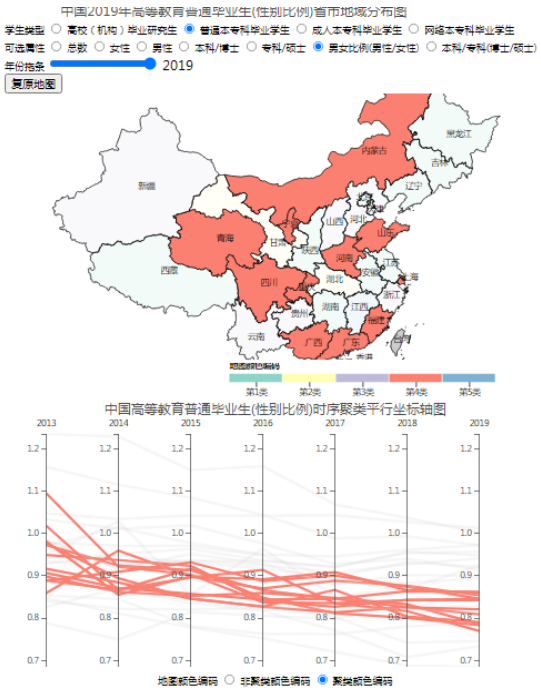

图 12 平行坐标轴图与地图的交互影响展示

4.5 地图与时间序列折线图的交互影响

通过点击地图具体省市，在右下图的详细视图中会展示对应省级行政区在某种高等教育类别下人口统计特征的详细情况。我们的设计严格遵循了 overview first, zoom and filter, then details on demand 的原则。

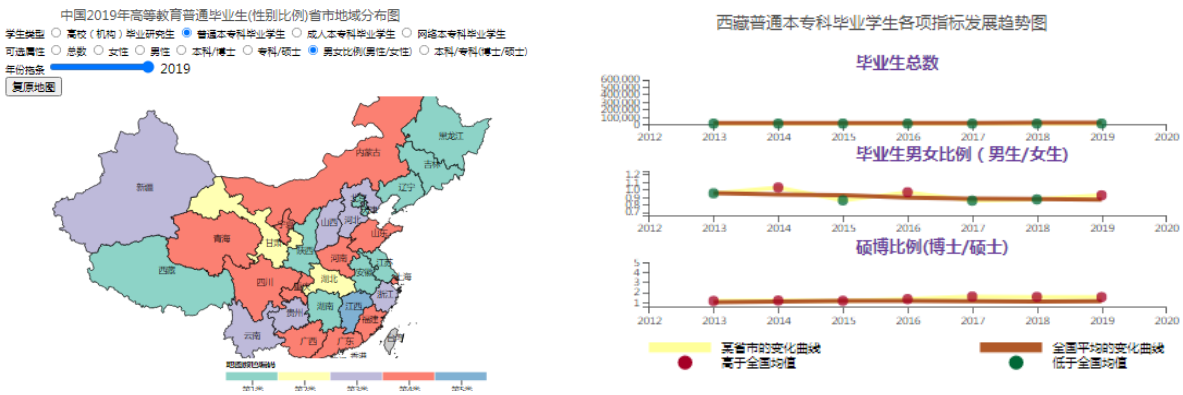

图 13 地图与时间序列折线图的点击交互展示

[i] http://www.moe.gov.cn/s78/A03/moe_560/jytjsj_2019/gd/