**数据对比工作方案（20160321）**

**说明：分析结果请根据下面划分的步骤分别建立各级文件夹。明确的步骤用“步骤一（二）”表示。**

|  |  |  |  |
| --- | --- | --- | --- |
| 步骤一、重名者 |  |  |  |
| 步骤二、“季”的差异问题 |  |  |  |
| 步骤三、数据库数据的匹配程度 | 步骤（一）、同一官职的任职次数和科举出身的变化 | 步骤1、公历年”出现的次数 |  |
| 步骤2、“科举”类型大于1 |  |
| 步骤（二）、不考虑公历年 | 步骤1、不考虑季节 |  |
| 步骤2、考虑季节 | (1) |
| (2) |
| (3) |
| (4) |
| (5) |
| 步骤（三）、数据库1、2与数据库3的比对 | 步骤1、数据库1与数据库3对比 | 步骤（1） |
| 步骤（2） |
| 步骤（3） |
| 步骤（4） |
| 步骤2、数据库2与数据库3的对比 | 步骤（1） |
| 步骤（2） |
| 步骤（3） |
| 步骤（4） |

**步骤一重名者**

分析前的数据处理：排除重名者：如果“官职”“姓名”相同的前提下，如果两条数据任职时间差别在50年以上，可以认定不是同一人，即为**重名者**，需要排除。

并将所有的重名者的百分比和数据输出。

**步骤二、“季”的差异问题**

1. 新的划分标准：

春季--农历1、2、3月；夏季：农历4、5、6月；秋季：7、8、9月；冬季：10、11、12月。

1. 拟要分析的结果：

**数据库1与数据库2的时间差统计，说明：以数据库1的“季”为基准，来计算数据库2与数据库1时间的差值。由于数据库1的时间只有“季”而没有“月份”，所以时间差以月为单位的计算方法为：“提前”是以数据库1季节的开始月开始计算；“延后”是以数据库1季节的结束月开始计算。例如，“提前一季”，即数据库2的时间比数据库1的时间提前一季。“提前1个月”，即数据库2的时间比数据库1的时间提前一个月，数据库2为“3月”，数据库1为“夏季”。**

示例：

|  |  |
| --- | --- |
| 时间差（单位季） | 时间差（单位月） |
| 提前三季 | 提前9个月 |
| 提前8个月 |
| 提前7个月 |
| 提前二季 | 提前6个月 |
| 提前5个月 |
| 提前4个月 |
| 提前一季 | 提前3个月 |
| 提前2个月 |
| 提前1个月 |
| 同一季 |  |
| 延后一季 | 延后1个月 |
| 延后2个月 |
| 延后3个月 |
| 延后二季 | 延后4个月 |
| 延后5个月 |
| 延后6个月 |
| 延后三季 | 延后7个月 |
| 延后8个月 |
| 延后9个月 |

具体计算方式是：如果月份为3，与夏季（4/5/6）比较，就是数据库2比数据库1提前1个月；如果月份是2，就是数据库2比数据库1提前2个月。同理，如果月份是7，就是数据库2比数据库1延后1个月；如果月份是8，就是数据库2比数据库1延后2个月。依次类推。（其中，**计算方法可以简单理解为：“提前”是以数据库1季节的开始月开始计算；“延后”是以数据库1季节的结束月开始计算。**）

如果月份是4或5或6月，那就是说明数据库2和数据库1在同一季中。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 季节 | 提前3个月 | 提前个月 | 提前1个月 | 夏季 | | | 延后1个月 | 延后2个月 | 延后3个月 |

以月为单位进行统计分析，输出各种分类标准下的频数和百分比，以及符合和不符合这种情况的数据。

**步骤三、数据库数据的匹配程度**

**步骤（一）同一官职的任职次数和科举出身的变化**

1. 步骤1在控制变量“姓名”“官职”的基础上，统计“公历年”出现的次数。即同一人任同一官职的次数。分别分析数据库2和1。输出数据结果。
2. 步骤2在控制变量“姓名”的基础上，统计“科举”类型大于1种出现的次数。只分析数据库1。（其中，包括“科举”为空的情况）并输出数据结果。

**步骤（二）不考虑公历年**，在修改过的数据库1和数据库2基础上的比较：

步骤1、不考虑季节，在官职+姓名基础上的，民族、旗分、科举的匹配情况（详见**☆**）。

步骤2、考虑季节：

**（1）季为空（数据库2）所占的百分比（不需要有☆的比较）**

**（2）同一季的比较**

**（3）连续二季，即提前或延后一季**

**（4）连续三季，即提前或延后二季**

**（5）相隔一年的匹配，即提前或延后三季**

**☆**在以上两个大前提的基础上，比较具体项如下

（1）比较**①**，官职、姓名的匹配情况。

（2）比较**②**，在①的基础上，“民族”的匹配情况。

（3）比较**③**，在①②的基础上，“旗分”的匹配情况。

（4）比较**④**，在①②③基础上，“科举”的匹配情况。

（5）比较**⑤**，在同一“官职”、“姓名”基础上，单独比较“科举”的匹配情况。

**④**科举

**③**旗分

**②**民族

**①**官职、姓名

**⑤**科举

官职、姓名

**注意：输出结果同时要输出各项的空值比例，使得各种结果相加为100%**

**步骤（三）数据库1、2与数据库3的比对。**

数据背景介绍：数据库1和数据库2的基本信息中“科举”一列有进士、举人、生员等等多种情况。而数据库3中默认所有人的“科举”都为进士。下图是数据库3的表头截图。在这部分分析中，只需用到“科年公历”和“姓名”两列的数据。



数据库1、2与数据库3的关系是部分姓名重叠、和部分科举信息重叠。

步骤1、数据库1与数据库3对比。

步骤（1）

前提：数据库1“科举”=“进士”或“翻译进士”

比较内容：数据库1中有“姓名”&数据库3无“姓名”

注：（1）的比较是针对科举信息重叠的人，来看数据库1或2的人（“科举”=“进士”或“翻译进士”）是否都在数据库3中。

步骤（2）

前提：数据库1和3有共同的“姓名”

比较内容：①数据库1“科举”=“进士”**或**“翻译进士”

②数据库1“科举”≠“进士”**和**“翻译进士”

注：（2）的比较是针对姓名重叠的人。首先，数据库1和3有共同的姓名说明两个数据库有相同的人。①②的比较实际是看这部分相同的人其“科举”的情况，一是“进士”或“翻译进士”即①； 二不是“进士”和“翻译进士”，即②。

然后对 不是“进士”和“翻译进士”，即②的这些数据进行（4）的分析。

步骤（3）数据库1和数据库3 姓名相同的那部分人的，分别列出两个数据库这些限制条件的内容和百分比：①姓名相同；②姓名、籍贯相同；③姓名、旗分相同。

也就是：对于数据库1， 姓名相同的人/数据库1的总人数， （姓名+籍贯）相同的人/数据库1的总人数， （姓名+旗分）相同的人/数据库1的总人数。

对于数据库3，同理。

步骤（4）在步骤（2）的②的基础上（即数据库1和3有共同的“姓名”，且数据库1中“科举”≠“进士”和“翻译进士”的人），计算满足条件的数据时间差。即数据库3的“公历年”与数据库1中最后出现的“公历年”的差值（说明：“最后一次”是针对数据库1中公历年有1个以上的情况，只计算最后一个公历年，也就是数值最大的那个；不是进士和翻译进士的人任职的公历年与数据库3公历年的时间差，数据库3减去数据库1的公历年。

输出“姓名”“数据库2时间”“数据库3时间”“时间差”。

3、输出的结果：

（1）“前提”条件下的数据为步骤（1）或（2）的总体（即100%），首先输出“前提”占总数据（数据库1）的百分比。

（2）匹配和不匹配的频数、百分比；数据。（此百分比是以前提下的数据为总体100%）

不仅列出数据库1的匹配（不匹配）数据，也将数据库3的匹配（不匹配）数据列出，以便于两个数据库的人工比对。

步骤2、数据库2与数据库3的对比。

步骤（1）

前提：数据库2“科举”=“进士”或“翻译进士”

比较内容：数据库2中有“姓名”&数据库3无“姓名”

注：（1）的比较是针对科举信息重叠的人，来看数据库2的人（“科举”=“进士”或“翻译进士”）是否都在数据库3中。

步骤（2）

前提：数据库2和3有共同的“姓名”

比较内容：①数据库2“科举”=“进士”**或**“翻译进士”

②数据库2“科举”≠“进士”**和**“翻译进士”

注：（2）的比较是针对姓名重叠的人。首先，数据库2和3有共同的姓名说明两个数据库有相同的人。①②的比较实际是看这部分相同的人其“科举”的情况，一是“进士”或“翻译进士”即①； 二不是“进士”和“翻译进士”，即②。

然后对 不是“进士”和“翻译进士”，即②的这些数据经行（4）的分析。

步骤（3）数据库2和数据库3 姓名相同的那部分人的，分别列出两个数据库这些限制条件的内容和百分比：①姓名相同；②姓名、籍贯相同；③姓名、旗分相同。

也就是：对于数据库2， 姓名相同的人/数据库2的总人数， （姓名+籍贯）相同的人/数据库2的总人数， （姓名+旗分）相同的人/数据库2的总人数。

对于数据库3，同理。

步骤（4）在步骤（2）的②的基础上（即数据库2和3有共同的“姓名”，且数据库2中“科举”≠“进士”和“翻译进士”的人），计算满足条件的数据时间差。即数据库3的“公历年”与数据库2中最后出现的“公历年”的差值（说明：“最后一次”是针对数据库1中公历年有1个以上的情况，只计算最后一个公历年，也就是数值最大的那个；不是进士和翻译进士的人任职的公历年与数据库3公历年的时间差，数据库3减去数据库2的公历年。

输出“姓名”“数据库2时间”“数据库3时间”“时间差”。

3、输出的结果：

（1）“前提”条件下的数据为步骤（1）或（2）的总体（即100%），首先输出“前提”占总数据（数据库1）的百分比。

（2）匹配和不匹配的频数、百分比；数据。（此百分比是以前提下的数据为总体100%）

不仅列出数据库2的匹配（不匹配）数据，也将数据库3的匹配（不匹配）数据列出，以便于两个数据库的人工比对。