结果分析:

一、 **季为空：**（？季为空的数据总数是多少，63 个？（姓名、官职匹配与不匹配的数量和））

结果：

季为空	匹配（个）	（%）	不匹配（个）	（%）
姓名、官职	9	1.11	54	6.67
姓名、官职、民族	7	0.87	56	6.92
姓名、官职、民族、旗分	5	0.62	58	7.17
姓名、官职、民族、旗分、科举	3	0.37	60	7.42

以数据库 1 为基础，数据量为 809 个。数据库 1 中季为空的数据一共 63 个，占 7.79%，所占比例不大。如果分析任职时间上两个数据库的异同，必须要考虑“季”的问题。在季为空的所有数据中，匹配的结果一共 9 个数据，其中共 5 人（其中有 4 人出现过两次），所以对总体的影响不大。

由表中可知，不匹配的比例远远高于匹配的比例，前者约是后者的 6-7 倍，这一结果与下面分季节比较的结果相异（匹配比例高于不匹配的比例），影响因素不清。

思考：应当增加一项比较，在不考虑季的影响下，只在官职、姓名的基础上比较民族、旗分、科举的匹配情况，以此结果作为基础，与考虑季的结果相比，才能得出季的影响程度有多大。即：

		匹配	不匹配
1 不考虑“季”		（比较 1、 2、 3、 4）	
2 考虑“季”	1） 季为空		
	2） 同一季		
	3） 连续两季		
	4） 连续三季		
	5） 一年		

Wang Yang 16-2-1 8:01 PM
批注 [1]: 这里的不考虑季，需要考虑“公历年”一样吗？

二、 同一季

同一季	匹配（个）	（%）	不匹配（个）	（%）
姓名、官职	663	81.95	83	10.26
姓名、官职、民族	449	55.50	297	36.71
姓名、官职、民族、旗分	377	46.60	369	45.61
姓名、官职、民族、旗分、科举	232	28.68	514	63.54

Wang Yang 16-2-1 8:06 PM
批注 [2]: 这个同一季，有个隐含条件就是公历年相同，但是加上这个条件，匹配的结果会少很多。

Wang Yang 16-2-1 8:07 PM
批注 [3]: 这个 663 应该是没考虑“公历年”相等这个限制条件，如果加上这个限制，匹配的会少很多，具体数据我明天发给你。

以数据库 1 为基础，数据量为 809 个。“同一季”意为官员任职时间上，两种记载体系中的时间为同一“朝代”、同一“季节”。其中，数据库 1 中有明确的“季节”

记载，数据库 2 的资料来源中，有具体的任职月份，与数据库 1 的资料来源相比，时间记录更加详细，据此可以确定记录的“季节”。其具体的分析结果如下：

1、 姓名、官职的匹配情况

由上表可知，同一年、同一季的姓名、官职匹配情况高达 82%，说明数据库 1 的资料来源与数据库 2 所用的精确的材料相比，时效性并不比差。原因可能是因为以“季”为单位做比较，时间跨度为 4 个月，在这一较长的时间段内职官任用的信息上传下达还是比较及时的。如果两个数据库按照“月”为单位进行比较的话，这一匹配率有可能会下降。（虽然数据库 1 不可能提供“月”的信息。）

仔细查看数据库 1 中不匹配的职官表，发现其中有许多官员的名字不完整，原因是资料录入过程中因为文字模糊或不宜辨认导致。其中，姓名中有“？”的有 14 人，占不匹配总数的 20%（14/83）；姓名中有字号的如“绵宜（佩卿）”有 4 人，占不匹配总数的 9.64%（4/83）。这种“姓名”上的不统一会导致结果的偏差。笔者认为，如果将缺失的姓名补充完整并将字号去掉再进行比较的话，匹配率会提高至少 1.5%。下面的比较匹配率也会上升。

Wang Yang 16-2-1 8:10 PM
批注 [4]: 针对这个问题，可以在处理之前人为手动地进行纠正。

建议：将姓名参考数据库 2，如果朝代、官职、季节能够大部分统一的话，可以将姓名统一（根据姓名字的个数、已知的文字），有字号的直接去掉。

Wang Yang 16-2-1 8:10 PM
批注 [5]: 这个建议可以接受。

2、 姓名、官职 、民族

数据库 1 中同年、同季任职的官员，其民族的匹配率为一半，匹配率较低。查看输出的不匹配之数据表发现，对结果有直接影响的是具有先赋身份的“宗室”与“觉罗”。在数据库 1 的资料来源中，具有宗室、觉罗身份的官员，其“民族”全部是“满洲”，共有 184 人，占不匹配官员数量的 61.95%。而在数据库 2 中，笔者发现，其中有 179 个“宗室”、“觉罗”身份的官员记载，而其中具有“满洲”身份的只有 18 人，占宗室、觉罗人数的 10.06%。可见，两个数据库的记录体系不同，数据库 1 来源的资料记载个人履历信息时包括政治身份、民族、旗分。而数据库 2 的记载体系对具有宗室觉罗身份的人来说，只记载其政治身份、旗分，民族似乎都是省略的，默认为满洲。

建议：将所有具有先赋身份（宗室、觉罗）的人，其民族属性补全，再进行比较，匹配率应当会提高 19%左右，当为 74%左右，比较合理。

Wang Yang 16-2-1 8:17 PM
批注 [6]: 如果数据库 2 中，民族为空的话是将数据库 2 中身份为“宗室、觉罗”的人的民族填成“满洲”吗？

3、 姓名、官职、民族、旗分

旗分的比较同民族的比较，具有先赋身份的人群对结果影响较大。民族、旗分的比较虽然有差异，但是差异不是很大（大约 9%）。

4、 姓名、官职、民族、旗分、科举

由上表可知，科举的匹配量大约 1/4 左右。由于在比较 2、 3 中，姓名、民族对结果有一定的影响，所以分析结果存在一定的偏差。其中，在比较 4 中输出的不匹配的数据表中，同时具有先赋身份和科举出身信息的人有 58 人，（去除比较 1 中姓名缺失的部分数据后）占不匹配人数的 11%，笔者推测，其中大部分人因为姓名、民族不匹配而被忽略。将以上因素考虑其中，科举的大概匹配率应在 35%左右。参考输出的不匹配数据表，其中具有科举信息的有 188 条，占不匹配人数的 36.58%，占数据库 1 总数的 23.24%，这一比例相当高。也就是说，数据库 1 中有 23.24%的具有科举出身的人被忽略了。由此，笔者推测，数据库 1 记载的科举信息会较数据库 2 详细。

建议：如果将姓名、民族信息补充完善后再进行比较，匹配率可能会上升。

思考：除此之外还有没有可以用来比对的有关科举信息的详细资料？进士录中所有旗人的信息形成数据库 4。目前还有数据库 3 可供对比研究。

数据库 3 中有较全的科举信息。可以分别以数据库 1 和 2 为基础，只比较“姓名”“出身”（姓名去重后）就可以，看两个数据库哪个科举信息更全面。

三、需进一步跟进的工作：

1、 需要连续 2 季、连续 3 季、1 年的匹配情况，据此与同一季的匹配情况进行对比。

2、 将数据库 1 中姓名具有“？”的数据和具有字号的数据根据数据库 2 校对，能确定、改正的进行修改。

3、 将数据库 2 中具有宗室、觉罗身份的官员的民族信息——满洲补充完整。

4、 针对“科举”状况，再进行一下比较，数据库 1、数据库 2 的姓名去重后，以人为单位，不重复，分别与数据库 3 中的“姓名”“入仕途径”比较，看两个数据库在“科举”信息上哪个比较准确。

数据库 3 中的“入仕途径（类别）”即科举信息。

5、 针对“季”的情况，考虑到“季为空”的不匹配数据比例较大，还需要加入另一项比较，即不考虑“季”的影响下，在“官职”、“姓名”的基础上进行民族、旗分、科举的比较。（前期需要数据处理工作，即根据官职、姓名两项进行去重

Wang Yang 16-2-1 8:19 PM
批注 [7]: 这里需要考虑“公历年”相等吗


数据库3 [只读] - Microsoft Excel
经典菜单
开始
插入
页面布局
公式
数据
审阅
视图
开发工具
ABBYY FineReader 12
自 Access
自网站
自文本
自其他来源
现有连接
获取外部数据
全部刷新
连接
属性
编辑链接
连接
排序...
筛选
清除
重新应用
高级
排序和筛选
分列
删除重复项
数据有效性
合并计算
假设分析
数据工具
C292
科举: 进士(笼统)

使得“官职+姓名”唯一化。）