

HQMS肾病数据分析

邱玮 郭嘉明

2018.03.23

Outline

- 项目介绍及研究目标
- 主要研究方法
- 可视化结果及分析

项目介绍及研究目标

- CKD & HQMS
 - 慢性肾脏病(chronic kidney disease, CKD)由于其患病率高、易发展为终末期肾脏病的特点，已成为全球性的公共卫生问题。
 - 医院质量监测系统(Hospital Quality Monitoring System, HQMS)涵盖了全国三级医疗机构海量的住院患者数据，包括了患者的一般信息、出院诊断以及就诊医院信息等，足以满足对于CKD疾病谱及患者跨区域的现状和影响因素分析。
- 目标
 - 了解我国CKD疾病谱的变迁情况;
 - 探索我国CKD患者异地就诊的现状影响因素，提出优化肾脏专科建设及资源配备的政策建议。

Outline

- 项目介绍及研究目标
- 主要研究方法
 - 对应分析
- 可视化结果及分析

主要方法——对应分析

- 对应分析的主要目的是寻求列联表行因素和列因素的基本分布特征和它们的最优列联表示。它通过构造一些指标来反映行和列之间的关系。这些指标同时告诉我们在一行里哪些列的权重更大以及在一列中哪些行的权重更大。
- 对应分析的主要思想是按照重要性的降序来构造指标，使得表格的主要信息能以更小的维度来报告。比如，如果只有两个因子，则结果就可以在一个二维图上显示，该图正反映了表格中行与列之间的关系，即在同一行中哪些列分类更重要或同一列中哪些行分类更重要。

主要方法——对应分析

对应分析的原始数据是一个 n 行 m 列的矩阵 $X = (x_{ij})_{n \times m}$,
其中 x_{ij} 表示行变量($i=1,2,\dots,n$)在列变量 $j(j=1,2,\dots,m)$ 上的值。

$$r_i = \sum_{j=1}^m x_{ij}, i = 1, 2, \dots, n$$

$$c_j = \sum_{i=1}^n x_{ij}, i = 1, 2, \dots, n$$

$$W_{ij} = \frac{x_{ij}}{\sqrt{r_i} \sqrt{c_j}}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$$

将原始数据矩阵 X 变换成 W

主要方法——对应分析

对应分析将列变量间相似矩阵表示为 $W^T W$ 的形式，

将行变量间相似性矩阵表示为 $W W^T$ 的形式

令 $H = W^T W, Q = W W^T$

分别从矩阵 H 和矩阵 Q 出发进行因子分析。

根据两种因子分析的对偶性，只作一种类型便可获得另一种类型的结果。

主要方法——对应分析

设从矩阵 H 出发进行因子分析的因子负载为 A^R ,从矩阵 Q 出发进行因子分析的因子负载为 A^Q .
则有

$$A^R A^{R^T} = W W^T$$

$$A^Q A^{Q^T} = W^T W$$

主要方法——对应分析

取因子空间维度为 p 维，则 $A_{m \times p}^R$ 给出了 p 维因子空间中 m 个列变量点的坐标，则 $A_{n \times p}^Q$ 给出了 p 维因子空间中 n 个行变量点的坐标.且这两个因子空间为同一个空间.

用一对因子为坐标轴，做成列变量和行变量点的散点图，可以反应列变量与列变量，行变量与行变量，列变量与行变量之间的关系. 在因子空间的投影点图上，变量点的相似性取决于他们在图中距离的远近，即一个行变量与相距最近的列变量的关系最为密切.

主要方法——对应分析

- 从HQMS数据库中提取时间（季度）-空间（省份）-病种的三维tensor，所存数据为某季度某省某病的患病人数
- 多病因——按照空间（省份）对tensor进行切片，并对各省份的病种-时间（季度）列联表进行对应分析
- 单病因——对人数tensor求比例，用某季度某省的某种病的人数除以该季度该省所有病种的总人数计算出新的tensor，并对新的tensor按照病种进行切片，对各病种的空间（省份）-时间（季度）列联表进行对应分析

Outline

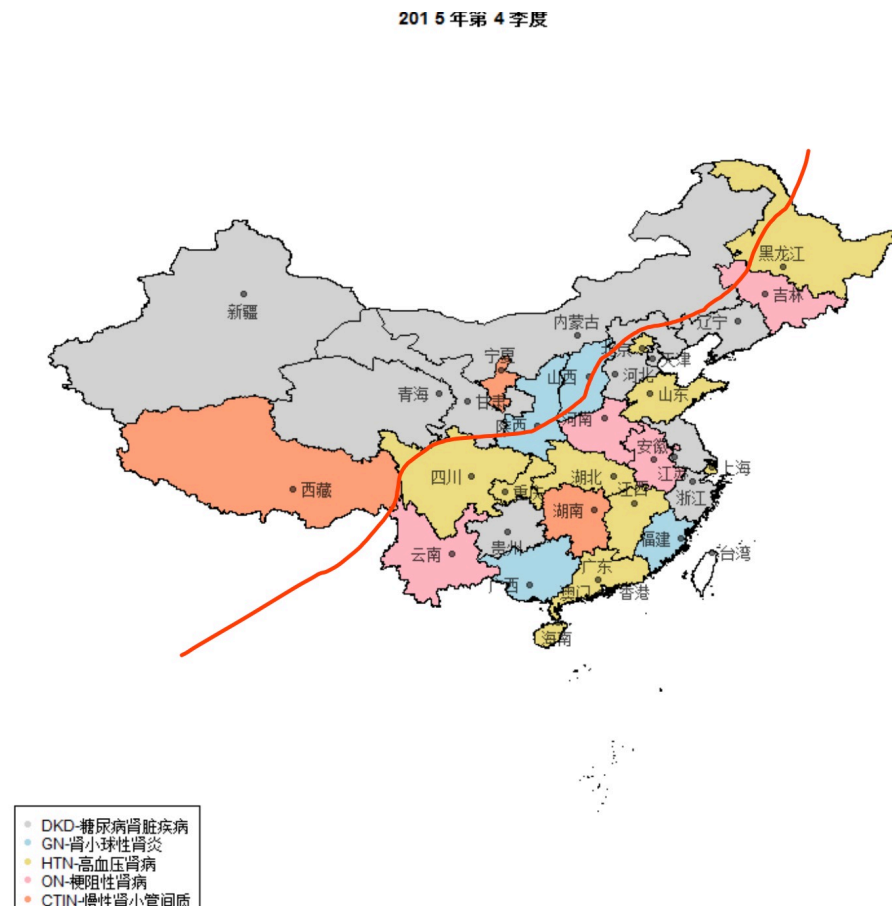
- 项目介绍及研究目标
- 主要研究方法
 - 对应分析
- 可视化结果及分析
 - 时间切片
 - 病种切片
 - 地区切片

可视化结果及分析

- 本文考虑的主要问题之一为CKD疾病谱的变迁情况，因此应着眼于如下几个变量包含的信息：**时间、空间、疾病种类**。
- 实际操作中，时间的单位取月或季度，空间的单位固定为中国的32个省级行政区，疾病种类固定6种，生成一个 $N \times 32 \times 6$ 的三维列联表 (tensor) 。
- 从该tensor的三个维度入手，切分成矩阵并做对应分析，可以有以下三种思路：按时间切片（分析省份和病种）、按病种切片（分析时间和省份）、按空间切片（分析时间和病种）。以下将分别讨论按三个维度分析的结果。

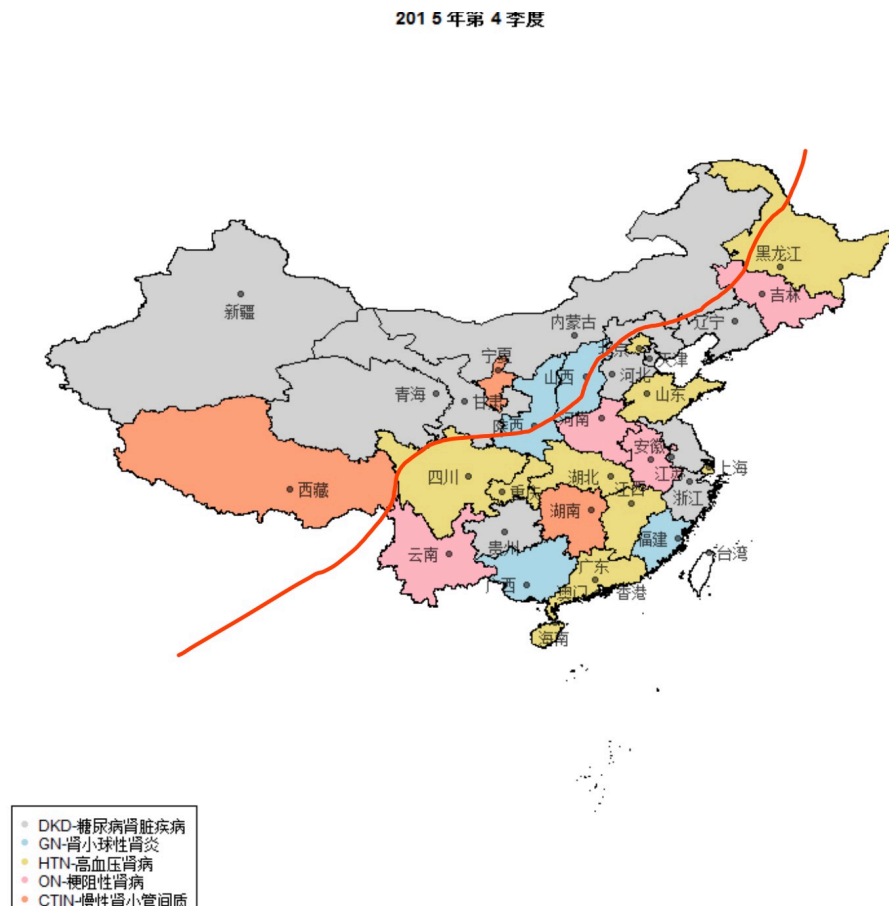
按时间切片

- 如图，按时间切片的结果中出现了一条明显的分割线。该分割线北起东北三省，跨越渤海至北京、山东、河南、湖北，南达重庆、四川。
- 此分界线以西、以北，呈现“由GN（蓝）、CTIN（红）过渡至DKD（灰）”的趋势；此分界线以东、以南，呈现“由各种病过渡至HTN（黄）、ON（粉）和DKD（灰）”的趋势。



按时间切片

- 此分界线两侧，疾病影响力发生突变的时间约为13年11月-14年2月。
- 一些特例
 - CTIN：西藏、宁夏、湖南三省的主要病因稳定在CTIN，未出现明显变化趋势。
 - GN：广西、福建、陕西三省的主要病因稳定在GN，未出现明显的变化趋势。



按病种切片

- **CTIN**

- 在2013年上半年，CTIN对宁夏、湖南、海南、辽宁、江苏、上海的影响较大。而到了2014年上半年，CTIN对青海、陕西、湖北、安徽、河南、江西、浙江的影响较大。到了2015年，CTIN对四川、陕西、河北、天津、广西、云南、福建的影响越来越突出，超过了CTIN对其他省份的影响。

- **DKD**

- 在2013年，DKD对湖南、浙江、青海、吉林、山东、广东的影响较大，到了2014年，DKD对东北多个省份的影响加大，对重庆、广西、江西的影响也较大，2015年，DKD对内蒙古、吉林、辽宁、河北、青海、重庆、河南、江苏、江西、广西的影响最大。

- **GN**

- GN影响较大的省份相对比较稳定，为青海、陕西、河南、河北、安徽、江苏、江西、内蒙古、广西。但是从地图的深浅可以看出，从2014年第一季度到2015年第四季度，GN对大多数省份的影响都在减小。

按病种切片

- HTN

- HTN影响较大的省份相对比较稳定，有内蒙古、吉林、辽宁、青海、四川、河南、江苏、四川、广西、江西，并且于2014年到2015年影响程度变化不大。

- ON

- 在2013年上半年，ON对新疆、湖南、宁夏、北京、上海的影响较大，之后到2013年下半年，ON对青海、吉林、陕西、山东、广东、上海的影响较大，之后在2014年，ON对不同省份的影响程度有一定的变化，并在2015年趋于稳定，2015年，ON对甘肃、四川、重庆、陕西、河北、河南、江西、广西的影响最大。

- Others

- 略

按地区切片

- 有些省份的DKD曲线有明显下降的趋势，即DKD对这些省的影响逐渐增大，如安徽、福建、甘肃、贵州、河南、黑龙江、江苏、辽宁、内蒙古、陕西、天津。
- 有些省份的GN曲线有明显上升的趋势，即GN对这些省的影响逐渐减小，如北京、广东、广西、海南、湖南、江西、青海、陕西、上海。
- 有些省份+有些病种的曲线非常类似，如广东、福建和浙江的ON曲线，广东和浙江的HTN曲线、DKD曲线、GN曲线，四川和重庆的GN曲线，四川和贵州的DKD曲线、HTN曲线、ON曲线，并且四川和甘肃的六种病的曲线形状都略有相似。

按地区切片

- 从ON、HTN、CTIN三种病的图像来看，也有一些省份的曲线有明显的上升或下降的趋势。
 - 如甘肃的CTIN的下降趋势、贵州CTIN的下降趋势、黑龙江CTIN的上升趋势、江苏CTIN的上升趋势、浙江CTIN的下降趋势
 - 山西HTN的下降趋势、上海HTN的下降趋势
 - 北京ON的下降趋势、贵州ON的下降趋势、河北ON的下降趋势、云南ON的下降趋势、浙江ON的下降趋势、河北ON的上升趋势。

感谢您的聆听与指导！