**基于人流数据的各地新冠疫情防控效果分析**

黄强盛

1. **摘要**

本文针对2020年初的我国新型冠状病毒肺炎疫情，通过全国各省/直辖市每日迁入人口规模数据和各地每日新增确诊人数数据，建立了模型，通过每日新增确诊人数和迁入人口风险系数数据的最佳拟合所使用的偏移量、时间窗口参数组合，作为外部迁入人流的管控、内部防控及内部人员安全意识的效果评估（每日全国各地都有外部/内部防控效果指标）。结果显示，上海等地内外防控相对比较好，黑龙江等地急需改进。并从理论、数据、新闻等多方面，印证了模型的准确性。最后，对全国各地复工顺序提出了建议。

关键词：新型冠状病毒肺炎;疫情防控;大数据分析;数据挖掘;数学建模

Analysis of the effect of new crown epidemic prevention and control based on the flow of people

Huang Qiangsheng

**Abstract**: This article aims at the outbreak of China's new type of coronavirus pneumonia in early 2020. Based on the daily immigration population data of provinces / municipalities across the country and the daily number of newly diagnosed patients in each country, a model has been established. The combination of the offset and time window parameters used for the best fit of the coefficient data is used as an effect evaluation of the external migration flow control, internal prevention and control, and internal personnel safety awareness (external / internal prevention and control are available daily throughout the country Performance indicators). The results show that the prevention and control of Shanghai and other places is relatively good, and Heilongjiang and other places need urgent improvement. The accuracy of the model has been confirmed from many aspects such as theory, data, and news. Finally, suggestions were made on the order of returning work across the country.

**Key words**: New Coronavirus Pneumonia; Epidemic Prevention and Control; Big Data Analysis; Data mining; Mathematical modeling

1. **背景**

自2019年12月，湖北武汉开始发生的新型冠状病毒肺炎疫情，已有2个月了，全国上下从一开始的不知情，到后来逐步重视外来人口排查和管控、关停部分人多密集场所、以及内部人员尽可能减少外出、出门带口罩、社区防范、等等。目前，除了湖北（病毒始发地）以外的地区，对病毒多少有恐惧心里，不清楚问题出在哪，不清楚各地疫情防控是否做得到位或者哪些需要改进的地方。本文通过所能搜集到的各地每日迁入人口规模数据和每日新增确诊人数，对全国各省、直辖市（港澳台因无法获取到人流数据没有进行分析，湖北因病毒始发地以及封城无过多人流数据的原因，也不做分析）建立的数学模型，可以从外部迁入人流的管控、内部防控及内部人员安全意识多方面作效果评估。

1. **模型使用的数据**

本文主要使用2类原始数据：各地每日新增确诊人数、各地每日迁入的人流规模（区分不同来源地区），并构造出第3类数据：各地每日迁入人流风险系数。本节对这些数据进行介绍说明。

本文将从百度迁徙获取到的“各地每日迁入人口规模数据”，简称为“人流”或“人流数据”（详见3.2节），基于此数据计算出的“各地每日迁入人口风险系数”简称为“人流风险”或“人流风险系数”（详见3.3节）。

* 1. **各地每日新增确诊人数**

自2020年1月11日起，全国各省、直辖市，陆续都有公布每日疫情新增确诊人数等数据，本文通过计算机爬取丁香医生网（<https://ncov.dxy.cn/>）每日疫情数据的各地累计确诊人数作为实时累计确诊人数，记录到本地数据库，因各地每日第一条公告都是公布截至到当天0点，累计确诊的人数，所以将采集到的数据中，各地每日第一条累计确诊人数有变化的值减去前一日第一条累计确诊人数有变化的值，作为各地前一日新增确诊人数。爬虫程序从2020年1月24日起开始持续执行，所以能准确地监控到累计确诊人的变化，并计算出每日新增人数，但1月24日之前的数据会有误差（某些地区该日期之前已有确诊人数，会把之前的累计确诊人数都算入到1月24日中去，虽然已尽可能手工修正了一些从1月17日起的数据，但不能确保都正确），而因这些1月24日之前的确诊人数相对后续疫情新增的人数较少，故本文还是从1月17日起进行数据分析。

* 1. **各地每日**迁入**的人口规模**

本次疫情，正好在春运及其返程期间，全国各地人口流动因素及其管控一定是影响疫情发展趋势的最重要因素，所以若能获取到足够的人流信息，一定能对疫情分析建模有至关重要的作用。通过百度迁徙（<https://qianxi.baidu.com/>）进行数据爬取（如图1所示中的今年北京每日迁入规模指数），获取到了每日全国省/直辖市迁入人口规模的数据，及其人口来源（按不同来源省/直辖市细分，如图2所示可查询到每日迁入的人口来源占比）。

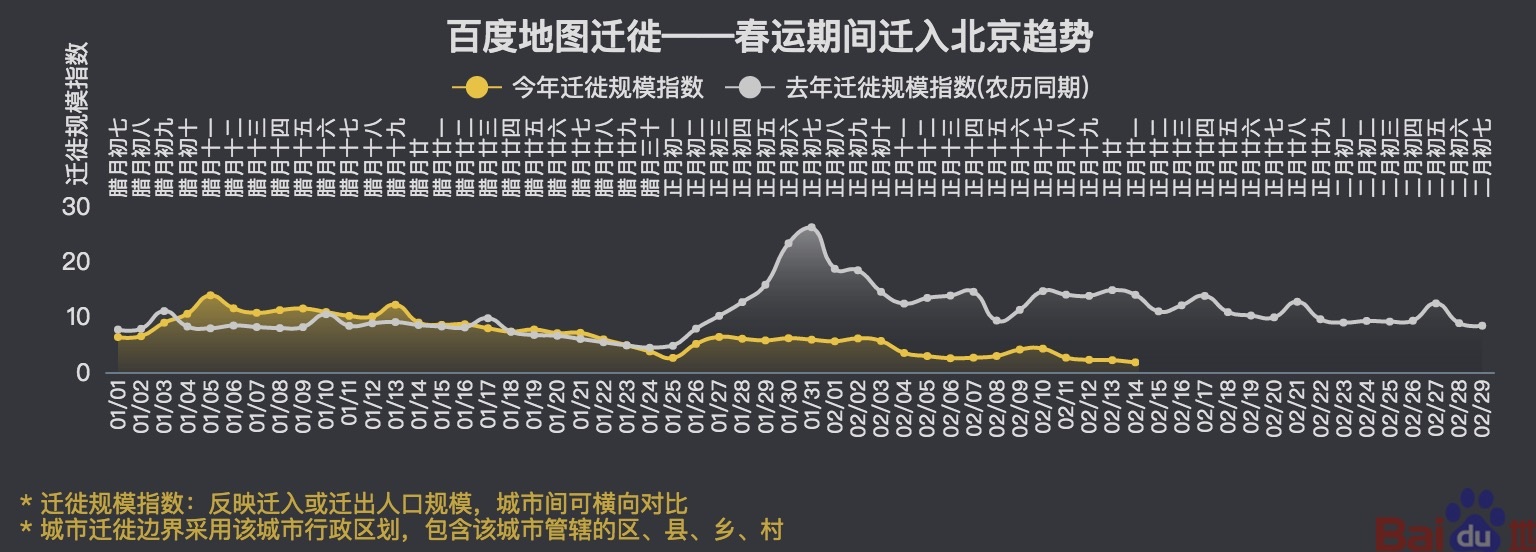


图1 百度迁徙查询到的北京今年每日迁入人口规模



图2 百度迁徙查询到的2020-02-14迁入北京的人口来源地比例

* 1. **各地每日**迁入**的人流风险系数**

因不同来源地的人流，其风险程度是不同的，比如在湖北封城前，从湖北来的和从其他地区来的相同规模的人流，对人流目的地区的疫情影响的风险是不同的。所以，各地人流数据，在输入到模型之前，再做一次如下方式的处理，即各来源地人流规模再乘以来源地近3日新增确诊人数，然后将各来源地的值相加，得到人流迁入地的每日人流风险系数：

人流迁入地每日人流风险系数 = Σ（来源地人流规模 \* 来源地近3日新增确诊人数）

公式中的“近3日”是这样由来的：通过测试发现使用近期的天数越少，相关性越大，判断是使用远期的确诊人数难以反映当前来源地风险的原因，但是考虑到单日新增人数的波动较大、潜伏期有几天的偏差，所以这里使用“近3日”。

另一方面，本文也考虑过在上述公式的括号中，再除以来源地的人口数，意味着相同确诊人数的来源地，人口数更少的话，病毒传播者迁入目的地的概率就大，但是经后续模型的相关性分析，发现相关系数并没有提升，但window、shift值（见下文）却有所增加才能拟合到差不多的相关系数，判断是因为流动人口相对非流动人口更容易感染病毒，所以最终使用如上公式。

1. **模型和数据分析**

本文的模型，通过每日新增确诊人数和迁入人流风险系数数据的最佳拟合所使用的偏移量、时间窗口参数组合，作为评估外部、内部疫情防控效果的评价指标。本节进行相应的理论分析和模型参数描述。

* 1. **相关系数计算**

模型使用的是3.1和3.3节描述的数据（后文分别简称为“每日新增”和“风险系数”），针对每个省/直辖市（后文简称为“地区”）从2020-01-17起的对这两组每日数据进行相关性分析。

分析发现，若直接对风险系数和每日新增数据计算相关系数，其值可能并不足够高，比如江苏省相关系数0.628（见图3）；甚至有几乎完全不相关的，比如黑龙江省相关系数仅-0.042（见图4）。（因风险系数值是个非常大的值，和每日新增人数画在一起，会令每日新增折线画成一条很低的横线，所以在画图时，对风险系数进行了调整，将风险系数最大值设为每日新增最大值，其余风险系数值等比例下调。）：

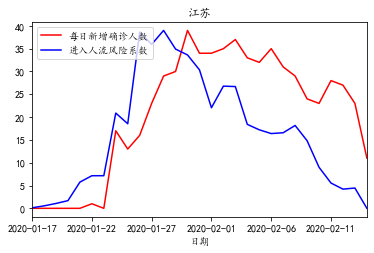


图3 shift=0,window=1的江苏数据，相关系数0.628

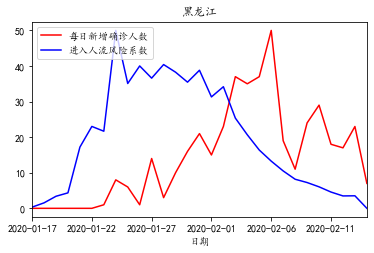


图4 shift=0,window=1的黑龙江数据，相关系数-0.042

* 1. **防控力度和偏移量、时间窗口的概念及其理论依据分析**

从疫情传播和防控方面，有如下两方面的考虑，相应地，引入了两个模型中的概念：

* + 1. **外部管控和偏移量（shift）**

一方面，对于外来人口的迁入，不同时期或不同地区，防范措施会不同，比如某些地区相对较早就开始严查迁入人流的体温等，这种严格的外部人流控制，会令外来感染者立即被确诊，所以风险系数高的当天或后一天，就有新增确诊人数的明显增加；反之，这些外部感染者最终被确诊的天数会来得相对更迟。因此，引入偏移量（shift）的概念：将风险系数偏移一定天数，比如3天，表示每一天的新增确诊和3天前的风险系数作比较。而4.1节中的计算，是在shift=0下的计算。

* + 1. **内部防范和时间窗口（window）**

另一方面，对于内部人员的管控和内部人员自身的防范意识，也会对疫情传播有影响，比如外部迁入该地区的感染人员（已经发病或潜伏期），并没有就医或居家隔离，内部人员不带口罩外出等防范意识差，会令外部风险延迟到后面几天才体现到每日新增确诊人数，即发生了在内部人传人的现象；反之，仅外部感染人员自己被感染了，只会有偏移量。所以，反应到模型中，引入时间窗口（window）的概念，比如时间窗口为10，表示连续10天的人流风险总和，会影响到第10天的确诊人数，从参考文献[1]中得知，95%置信区间的潜伏期是4.1-7.0天，所以如果是10天前迁入的感染者还能影响到内部的话，应该是内部人传人比较严重。而4.1节中的计算，是在window=1下的计算。

* 1. **使用偏移量、时间窗口处理风险系数并计算相关系数**

基于4.2节的分析和引入的概念，把4.1节的江苏、黑龙江分别取shift=0，window=8和shift=4，window=9，计算相关系数分别为0.952和0.865，如图5、图6，相关系数明显提高，2条折线图明显更拟合。

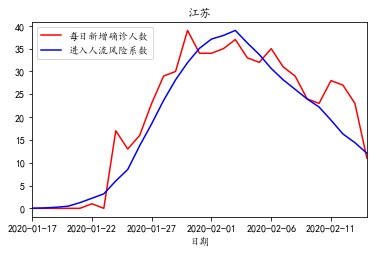


图5 shift=0,window=8的江苏数据，相关系数0.952

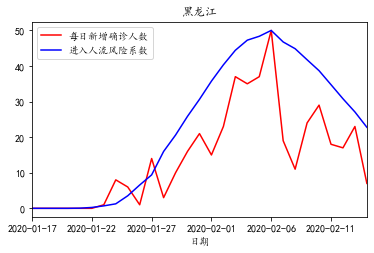
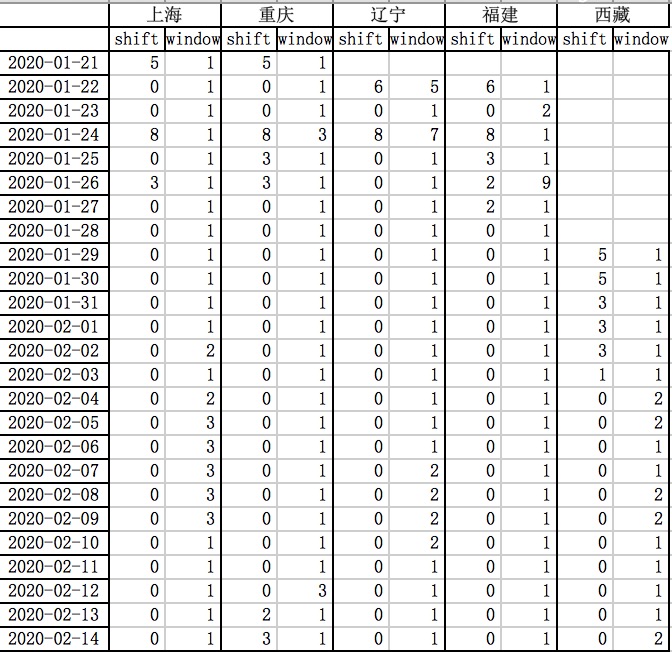
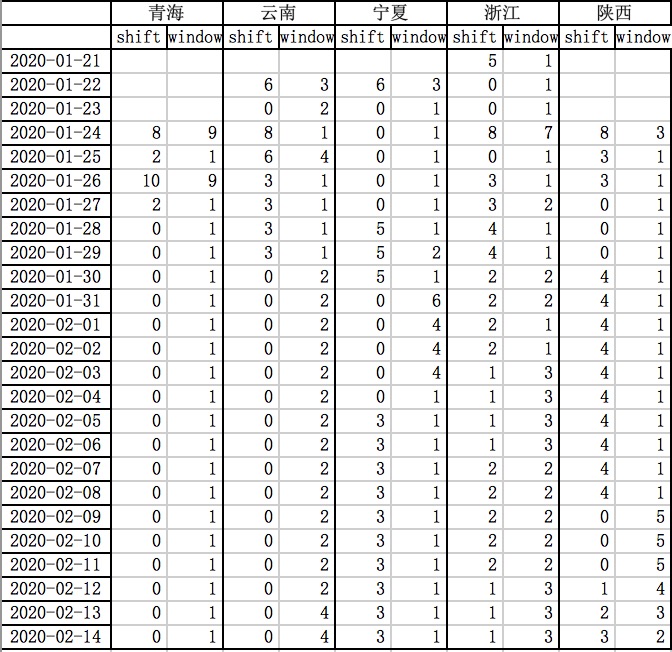


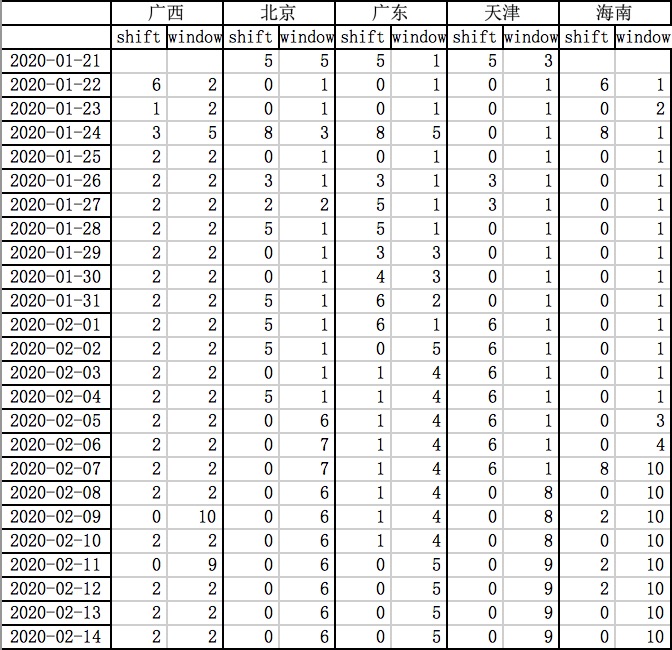
图6 shift=4,window=9的黑龙江数据，相关系数0.865

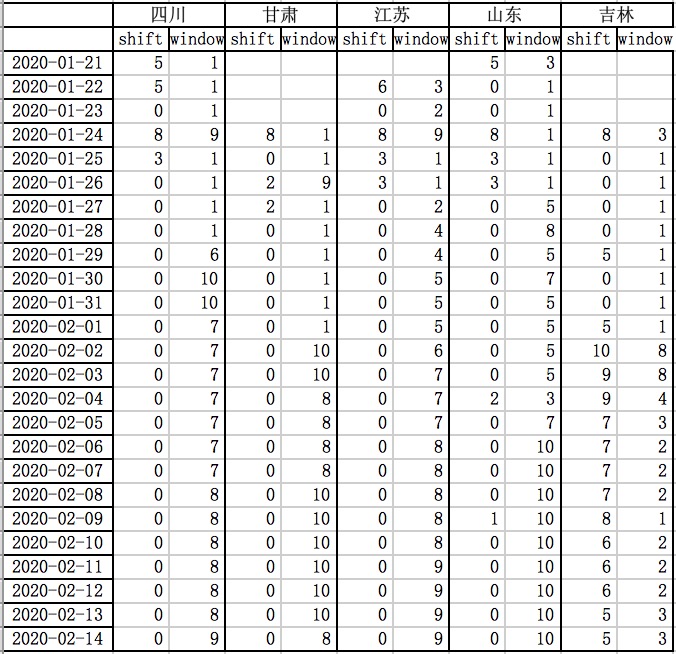
* 1. **各地每日疫情防控效果评估**

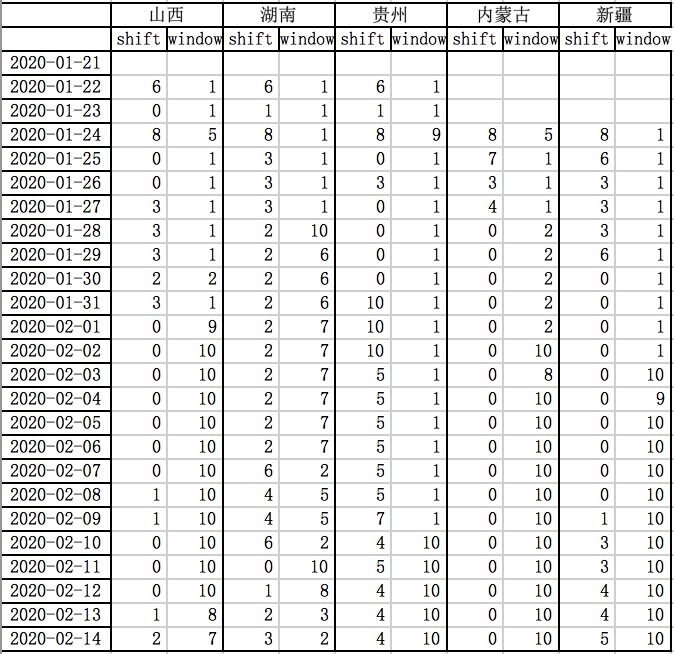
对所有地区，每天从2020-01-17开始，对偏移量shift取0-10，时间窗口window取1-10，每个地区排列组合共110组不同的shift、window进行处理后，再和每日新增数据进行相关系数计算，最终取最大相关系数对应的一组shift、window，整理得到如下表格，该表可以作为各地每日疫情内外防控效果的评估（表格中的地区按shift、window均值从小到大顺序排序，同时也可以作为各地复工安排（见第6节）的顺序，空白处表示当日该地区无新增人数，各地从首个有新增确诊人数的日期开始有计算结果）：











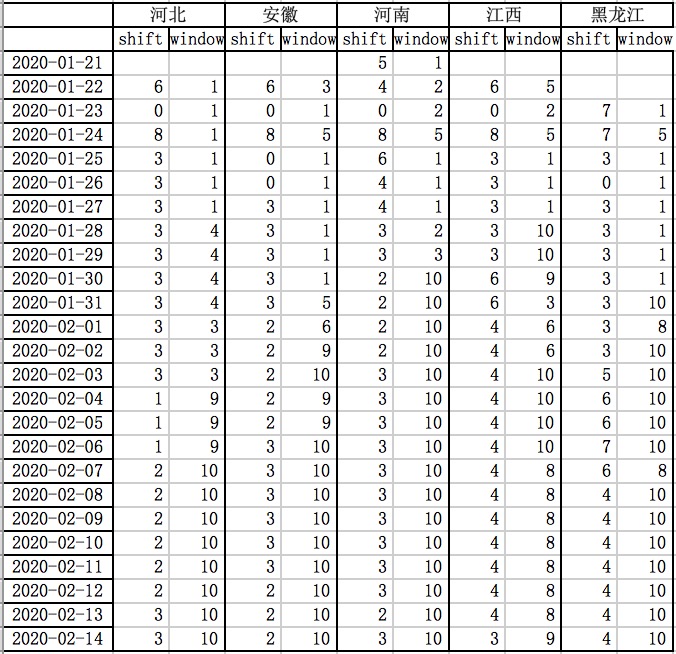


表1 模型输出的全国各地防控评估效果

从4.2节的分析可知：shift值大表示外部人流检查和管控相对较松，反之严格；window值越大表示内部防范和安全意识相对较弱，反之较强。

比如，从各地shift、window值来看：

上海等地，内外防控都做得很有效（shift、window值都较小）；

江苏等地，虽然外部迁入的人流管控做得很到位，但内部防控或内部人员安全意识需要加强（shift值较小、window值有些大）；

吉林等地，虽然内部防控或内部人员安全意识较强，但外部迁入的人流管控需要加强（shift值有些大、window值较小）；

黑龙江等地，内外防控都需要加强（shift、window值都较大）。

表格能作为对全国各地外部和内部疫情防范成效的概览，目前从总体上看，全国多数地区的外来人员管控普遍呈不断加强的趋势（偏移量shift值一开始较大，后来随着日期往后，逐步减少），而当前内部人员防控在很多地区需要越来越重视（window值随着日期往后，逐步增加）。

1. **相关新闻和数据对模型印证**

目前，丁香医生网公布的全国各地的累计确诊人数数据中，区分外来人员人数的有上海、北京、天津三地，截至2020-02-15下午，这三地的累计确诊外来数/总数分别是109/326、26/375、6/121，即上海内部管控较好（内部人员感染率较低），北京其次，天津相对较差，从这3地的最近window值来看，也是上海最低（近期值为1），北京次之（近期值为6），天津较高（近期值为9）。

另外，在2月8日，参考文献[3]（如下图7的部分网页截图），指出黑龙江对疫情不够重视、防范意识较差。从黑龙江的shift和window值上来看，也确实如此。

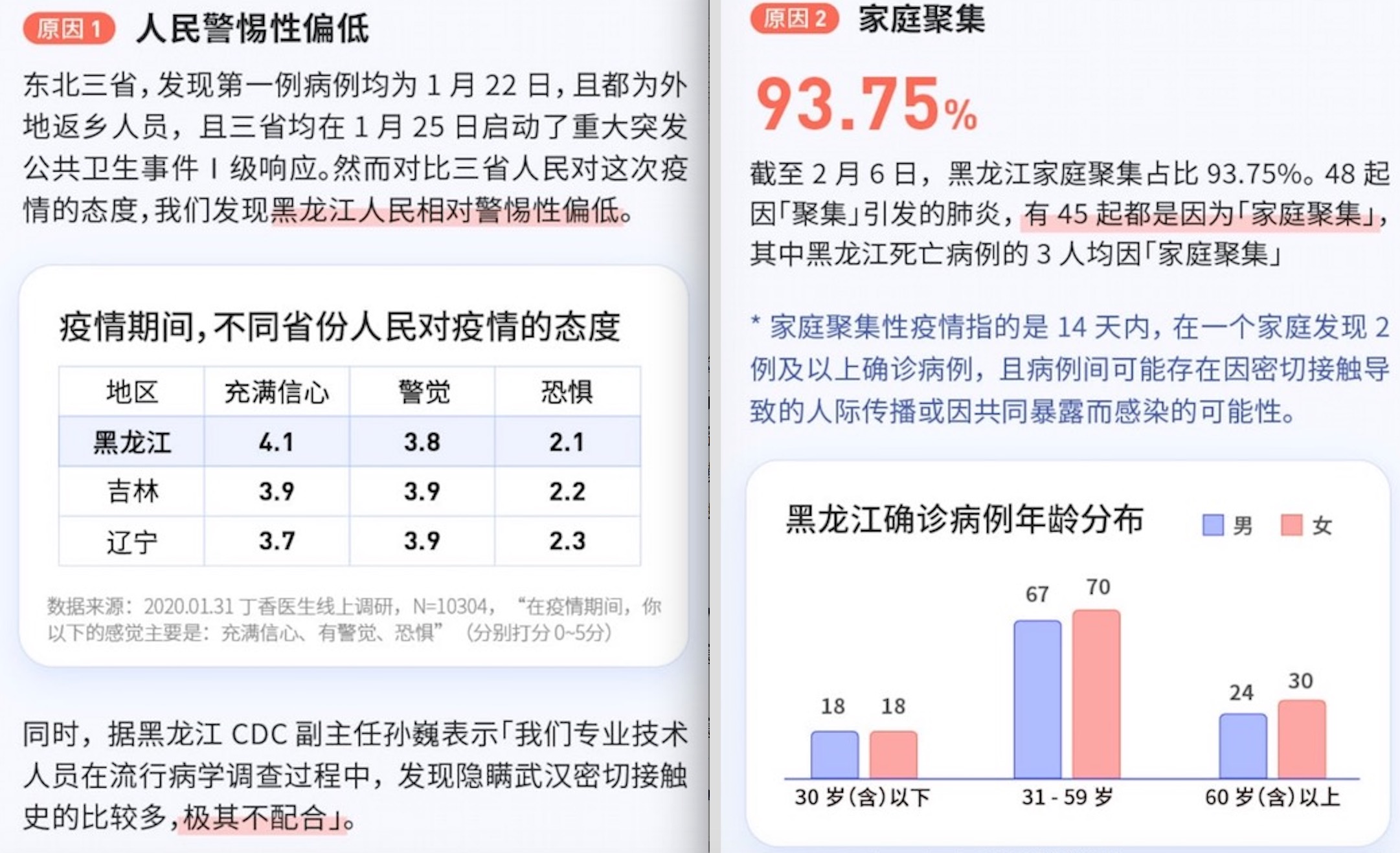


图7 丁香医生网分析黑龙江疫情严重的原因

并且，参考文献[2]的作者搜集到了上海、浙江、江苏、安徽、河南、四川6地（除去本文没有统计的湖北、武汉、全国）的R0数据，R0值从小到大排列为：上海、浙江、河南、四川、江苏、安徽，与表1中的排名几乎一致，仅河南有差别，表1中6地防控效果从好到差为：上海、浙江、四川、江苏、安徽、河南。猜测可能是个别地区临床R0指标统计上有出入或者其他原因导致有此差异。

1. **对全国各地复工的建议**

根据表1的结论，以及参考文献[2]中的R0数据，本文认为上海、重庆、辽宁、福建、西藏、青海这些疫情防控做得较好的地区，可以在确保R0小于1的前提下，先进行有序的复工，紧随后面排名的地区也可按表1中的排名顺序逐步开展复工，排名靠后的地区需要先加强防控力度，暂时还不能大规模复工（一旦复工规模大，R0大于1的可能性极大）。

1. **小结**

本文对新冠病毒肺炎每日新增确诊人数和人流数据进行分析，提出了每日迁入地区人流风险系数的概念和计算方法，以及使用偏移量、时间窗口的概念拟合数据并用于评估外部、内部疫情防控的效果。最后对个别防控做得好的地区，提出了有序复工的建议，以及全国各地复工的顺序。目前的疫情防控形势依然不能掉以轻心，相信我们在政府的领导下，全国人民共同的努力下，最终一定能打赢这场疫情防控阻击战！

1. **参考文献**

[1] Qun Li, M.Med., Xuhua Guan, Ph.D., Peng Wu, Ph.D., Xiaoye Wang, M.P.H., Lei Zhou, M.Med., Yeqing Tong, Ph.D., Ruiqi Ren, M.Med., Kathy S.M. Leung, Ph.D., Eric H.Y. Lau, Ph.D., Jessica Y. Wong, Ph.D., Xuesen Xing, Ph.D., Nijuan Xiang, M.Med., et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia[J]. The New England Journal of Medicine, January 29, 2020 DOI: 10.1056/NEJMoa2001316，[EB/OL].(2020-01-29).<https://www.nejm.org/doi/full/10.1056/NEJMoa2001316?query=featured_home>.

[2] 戴庆祝.疫情何时结束？靠谱的答案在这里！.aliren区块链联盟[EB/OL] .(2020-02-11).<https://mp.weixin.qq.com/s/6I1onP3jt3G76IYjy9YWlw>

[3]黑龙江确诊数为何居东三省最高. 丁香医生网[EB/OL].(2020-02-08) .<https://mama.dxy.com/outbreak/daily-of-nationwide-new?index=20200208&locationIds=999&share=true&entry=Previous_Review>