大数据信息采集及其偏差补救方法
——以甜党和咸党的口味地盘之争为例

苏毓淞　姚雨凌

摘　要: 在大数据的时代下,互联网虽然可以在很短的时间为舆情研究者提供海量的数据,但是,互联网获取的数据样本并非随机抽样,数据代表性的问题,使得研究者对这类数据的可靠性存在质疑。使用多层次回归和事后加权方法,调整互联网获取的数据,可以得到更合理的舆情估计值,重新分析甜党和咸党粽子口味地盘大战的例子中,结合从新浪微博自动抓取数据、分析文本的技术,实现特定议题舆情的自动采集,并提供回归调整的完整过程。本研究方法具有普适性,可以推广到其他的舆情主题。

关键词: 多层次回归;　事后加权;　代表性缺失;　大数据

基金项目: 清华大学数据科学研究院资助项目

作者简介: 苏毓淞,清华大学政治学系副教授(北京　100084) ; 姚雨凌,清华大学理学院数学科学系

DOI:10.13613/j.cnki.qhdz.002347

一、甜党和咸党的地盘之争

甜党(偏好吃甜食的网民) 和咸党(偏好吃咸食的网民) 在互联网平台上的战争,俨然成为网民针对特定议题出现两个极端意见的典型案例; 从 2013 年的豆腐脑口味之争,到 2014 年的粽子口味大战,大量的网民在数月的时间内,积极参与互联网平台的投票并发表意见,结果原来为了解中国民众口味地理分布的调查,却演变成为不同口味支持群体的地盘争夺战,最后获得的口味地理分布,由于样本代表性缺失的问题,反而不能反映真实的现状。

大数据统计在 2013 年成功反映了中国民众对豆腐脑口味呈现南甜北咸的地理分布。2013 年弹幕视频网(acfun. tv) 首次举办了中国网民对豆腐脑口味偏好的投票,数个月的时间,涌入了超过 6 万网民参与投票并发表意见,结果以地理分布来看,除了湖北、湖南、江西、福建、广东、广西、云南、海南、西藏、台港澳地区偏好吃甜豆腐脑外,其余地区都偏好吃咸豆腐脑,口味地图大致呈现北方爱吃咸豆腐脑,而南方爱吃甜豆腐脑的态势。以地理分布的总面积和地区总数上来看,咸党大胜甜党。① 虽然这些投票的网民并不能完全代表全体中国民众,但是这 6 万网民的投票结果,大致符合现状,体现了大数据的力量。

大数据统计在 2014 年却无法复制 2013 年成功的奇迹。2014 年弹幕视频网再度举办类似活动,针对粽子的口味进行投票。为了提高参与点击率,弹幕视频网将原先 1 人 1 票的规则改成每人每 4 小时可投 1 票,最后累计投票数果然大幅提升,突破 10 万人次,但是统计结果却呈现大幅度的偏差。

活动最开始 1 个月,口味分布呈现北方偏好吃甜粽子,而南方偏好吃咸粽子,地理分布符合中国

① 弹幕视频网于 2013 年 6 月 27 日上线发布《AcFun 投票活动: 你吃什么味的豆腐脑?》,http: //www. acfun. tv/a/ac715346,本文撰写所根据的数据是 2014 年 10 月 2 日通过该网页获得的结果,截至本文 2015 年 1 月 18 日投稿时,网站已清零所有数据。

各地民众吃粽子的现况,但是这一阶段咸党的分布,不论在总面积或地区总数上,远远落后于甜党,激发咸党发动反攻。接下来几个月,咸党透过刷票的手法,逐渐斩获中原以外的省份,然后东北落入咸党版图,最后中原也成为咸党的势力范围,统计结果是全中国的民众都偏好吃咸粽子,这显然与实际情形不符。① 这一案例也再次凸显了大数据遭遇代表性缺失问题时,很可能得出谬误的结论。

对此,本文将提出一种可靠的补救方案,通过多层次回归和事后加权的方法,调整互联网获取的海量数据,在一定程度上解决数据缺乏代表性所造成的信息偏差问题。

二、大数据信息偏差问题

随着大数据时代的来临,互联网上的海量数据逐渐进入专业学术研究者的视野。尽管社会科学领域的研究仍然以传统的实验设计和问卷调查等方式为主,然而成本低廉、数据量更大的互联网"大数据",日渐成为社会科学研究和舆情调查中不可忽视的数据来源。

"大数据"最常产生的误区是:人们认为,当样本量足够大的时候,便可以透过大数定律保证获得总体均值、方差等统计量的无偏估计,据此便可进行可靠的统计推论。但实际的情况是"大数据"存在无法周全的"信号问题"(邱东,2014:16—22);大部分的互联网平台采集的数据,其样本不具有总体代表性,信息是有偏差的;活跃在网络的群体,在特定议题上的意见,与线下的民众有一定的差距。

但是,不可否认的是,互联网数据在应对舆情的多样性和多变性方面有绝对的优势,这是传统的舆情采集方式所不能相比的,囿于经费、人力等原因,传统的舆情采集方式往往无法在时效性和样本量上满足大多数研究者的需求。因此问题不是要不要使用互联网数据,而是如何转变既有的数据分析思维,利用合适的分析工具正确使用它(朱建平 等,2014:10—17)。

本文认为,利用互联网数据的实时信息,配合适当的统计分析方法,进行舆情分析和预测,是大数据舆情分析首要突破的重点。这个攻关的工具就是多层次回归和事后加权方法(Multilevel Regression and Post-stratification,MRP)。哥伦比亚大学统计学系教授 Andrew Gelman 博士原先开发 MRP 方法的目的,是为了解决数据经过大量分层后,各个分组样本量不足,造成估计偏差的问题。Ghitza 和 Gelman(2013:762 – 776)运用 MRP 方法分析美国"全国选举调查数据"(National Election Survey,NES),实现了利用全局信息对分组的小样本进行调整,成功估算了美国总统大选的投票率。MRP 方法随即被 Wang 等(2014)应用,分析游戏平台返回的舆情调查数据,成功估算出美国总统大选中民众投票的预测值。简单来看,MRP 分为两个分析步骤:第一,使用多层次回归模型,依不同的人口和地理特征,对数据重新分层估算;第二,使用普查数据为权重,对新的分层数据进行重新加权。在重新打散和重新组合数据的步骤下,研究者一方面可以从不同层面观测估算的舆情,另一方面分层估算舆情再经过重新加权后,保证了分层估算的样本结构与总体结构一致,合理的提高估计效果(金勇进和张喆,2014:79—84)。因此,本文将结合多层次回归和事后加权方法,来调整互联网采集数据样本偏差的问题。我们使用的数据来源采集自新浪微博,重新分析咸党和甜党在粽子口味偏好的地理分布。作为中国最有影响力的网络社交平台之一,新浪微博的日均活跃用户达到千万量级。我们使用计算机自动地从新浪微博抽取语料数据,并藉由 MRP 方法进行调整,最终获得全中国层面合理的舆情估计。从不具有代表性的互联网数据出发,使用 MRP 方法调整获得新的估计值,利用地图可视化数据,直观看到舆情在各地区的分布。

① 弹幕视频网于 2014 年 5 月 29 日上线发布《AcFun 第二次甜咸大战:端午节投票活动》,http://www.acfun.tv/a/ac1199883,本文撰写所根据的数据是 2014 年 10 月 2 日通过该网页获得的结果,截至本文 2015 年 1 月 18 日投稿时,网站已清零所有数据。

三、微博数据采集与语义分析

中国居民对粽子的甜咸口味的偏好是一个简单易懂,且民众态度明显的议题。本文仅以此为例,展示完整的操作与分析过程,MRP 方法也同样适用于其他的议题。

利用 R 软件(R Core Team,2014) 中的 Rweibo 包(Li 和 Chen,2014) ,实现对新浪微博网页数据的搜索,得到 2013 年至 2014 年内 24 条热点极高(至少要有 500 条以上的回帖) ,讨论粽子甜咸话题的微博,然后在耗时不到 4 小时的时间,自动抓取了这些微博下面的评论文本,共计 20 765 条。从严格意义上来说,2 万多条的微博,算不上是"大"数据,但是比较来看,一般在国内具有全国代表性的调查数据,样本量也只不过在 4 000 到 10 000 个样本不等,调查时间则往往需耗时半年以上。因此,能在如此短的时间,获得 2 万余个样本,数据已经非常庞大。

微博的字数限制为 140 个字,因此更便于使用计算机进行文本语义分析。由于使用人工判读如此大量的文本缺乏效率,我们使用 R 软件进行文本分析。一个简单的判读准则是: 若该文本含有表达中立感情的中立语词 D(都喜欢、都可以、都爱、都吃) ,则刨去该中立文本; 然后在剩下的文本中,比较分析其中包含的甜语词 A(甜、糖、酱、枣) 和咸语词 B(咸、盐、肉、辣) 数量的差异,并用否定语词 C(不、否、没、难吃、差、讨厌、无法、无能) 进行转意,最后得到民众对于粽子甜咸偏好的数值,是为"甜味指数"sweet。

具体来说,假定用户 i 的文本为 w_i,则: 甜语词数目: $a_i = \sum_{x \in A} 1(x \in w_i)$; 咸语词数目: $b_i = \sum_{x \in A} 1(x \in w_i)$; 转折词数目: $c_i = \sum_{x \in A} 1(x \in w_i)$,该用户的"甜味指数"可以写作: $\text{sweet}_i \equiv (a_i - b_i) \times (-1)^{c_i}$,$\text{sweet}_i > 0$ 表示该用户偏好甜粽,$\text{sweet}_i < 0$ 表示偏好咸粽,$\text{sweet}_i = 0$ 则为无特定偏好。为检验上述语义判定原则的效度,我们随机抽取 100 条文本进行人工判断,发现正确率达 91%。

依照新浪微博开放的公开接口,可以在抓取微博文本的同时,抓取该用户的注册地域、性别等其他个人信息。因此,在 20 765 笔的微博数据中,包含用户性别、注册地点、发送设备、回帖内容,以及使用文本语义分析得出用户的甜味指数 sweet_i。

四、数据样本代表性分析

网络发言者绝不能当作全国人口的随机抽样。以地域、发送设备和性别计,获得的数据样本和全国人口的偏差如图 1 至图 3 所示。

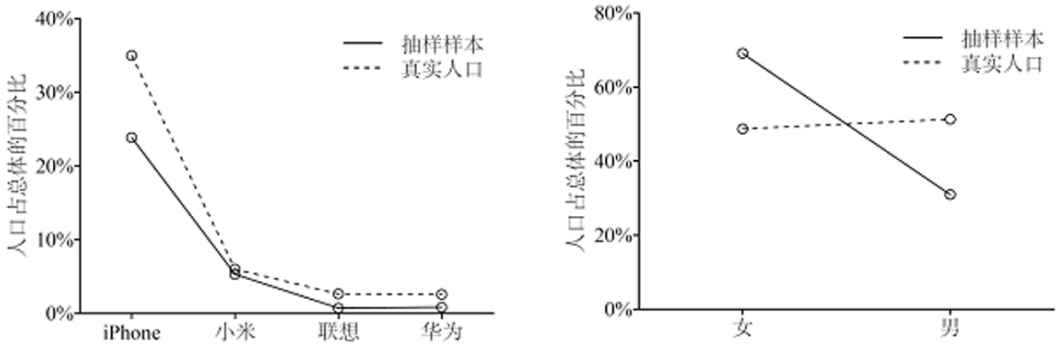

图1 客户端类型的比例与全国智能手机市场份额的偏差(左图) 和男女人数与中国总体人口比例的偏差(右图)

注: 左图中的虚线代表的中国智能手机市场份额,数据来源自市场调研机构 Flurry 2013 年 6 月的调查,右图中人口男女比例数据来源自第六次全国人口普查。

由于新浪微博公开接口的权限设置,无法获得用户的全部信息,但即便仅从上面展示的这些公开变量来看,我们也可以看到抽样样本和全国真实人口比例存在着偏差,因此这样的样本是不具有代表性的。尤其是在男女比例、各省人口分布、发送客户端上,明显看出样本中女性和发达地区用户比例过高,用户使用的四类发送客户端占比过低的问题。

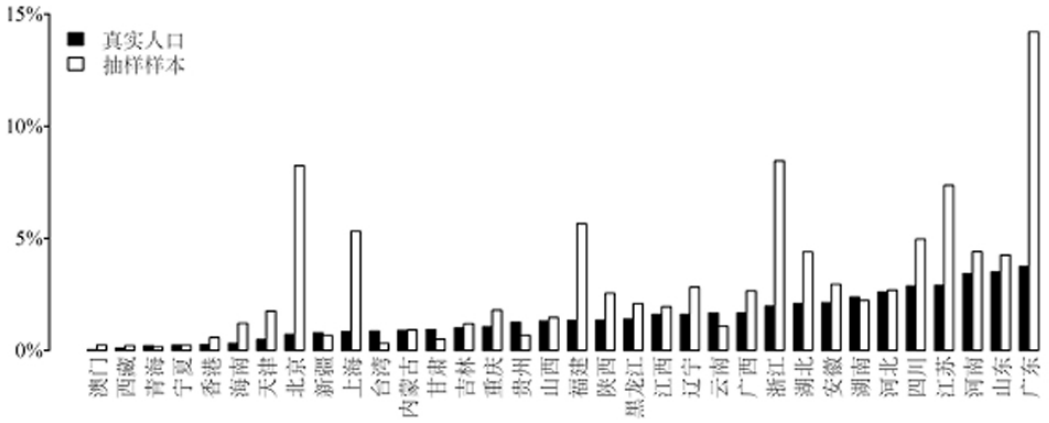

图 2　各省人数分布比例与中国总体人口比例的偏差

注: 图中中国人口各省比例数据来源自第六次全国人口普查。省份以普查人口数,由上而下排序。

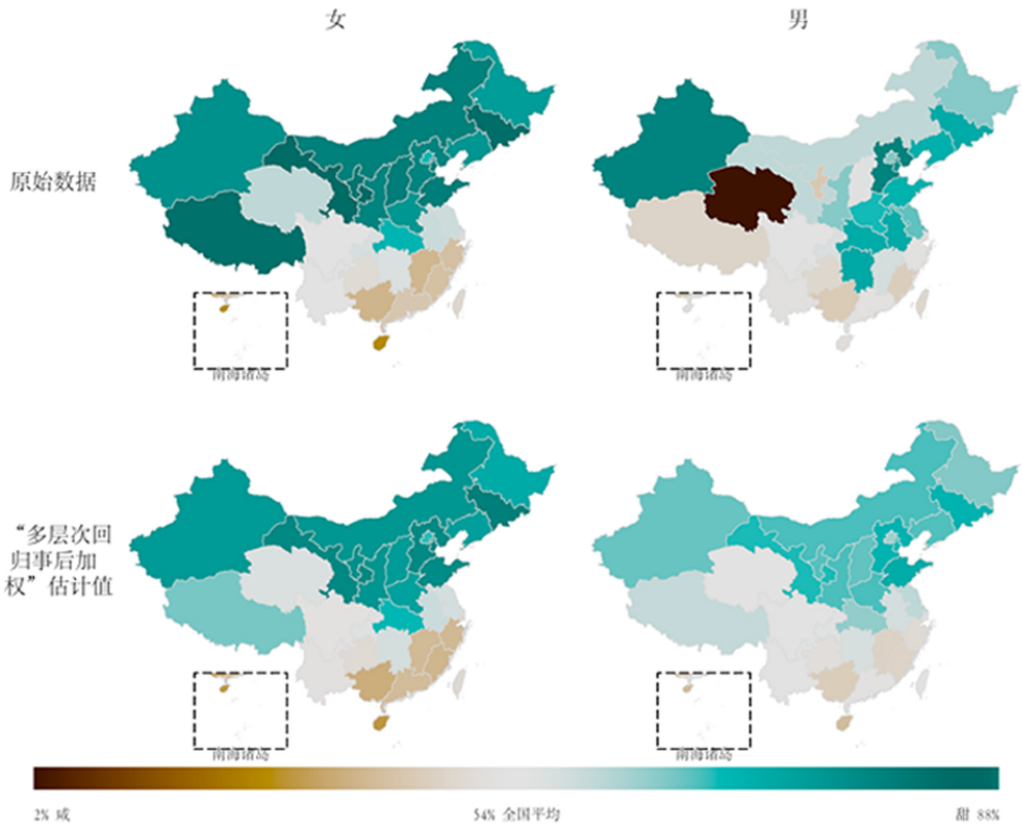

图 3　中国民众对甜粽子的偏好比例的地理分布图

注: 上两图为原始数据中各分层偏好甜的频率,下两图为使用 MRP 后拟合得到的各组对甜的偏好概率。

在不使用 MRP 方法的情况下,依照性别和 34 个省份地区切分数据,①绘制呈现民众对粽子甜咸偏好的地图(如图 3 第一行两幅地图),总的来看,全国的甜味指数均值为 0.49,说明中国民众偏好甜粽的人数稍微少于偏好咸粽的人数。不过,两幅地图呈现不少的噪声和错误信息。首先,在样本量比较大的女性用户地图中,呈现了民众对粽子口味偏好南咸北甜的分布,这与中国南方吃咸粽,北方吃甜粽的现状非常吻合,说明了网络大数据在一定程度上,仍然贴近现实,不过,在西藏地区,民众对于甜粽的偏好竟然高居全国首位,这可能是样本量偏小所造成的误差(参见图 2 西藏地区样本量)。其次,在样本量比较小的男性用户地图中,噪声和错误就更多了,个别北方省份呈现对咸粽的偏好,因此呈现不出南咸北甜的分布,其中,山西和宁夏地区偏好咸粽是明显的错误,通过采访数位来自山西和宁夏的同学,我们得知,在当地市场上几乎看不到卖咸粽的商家,并且家里从小吃的粽子,都是甜粽。至于青海地区对咸粽的偏好高居中国首位,更反映出因为组内样本量不足的关系,出现极端的估计值。

五、多层次回归和事后加权

为了解决上述样本代表性问题以及分组后样本点过小的问题,使用 MRP 方法进行调整,沿用 Gelman 等(2013)的记号如下:

把第 i 个样本,按照性别、地区进行分组,记为 $j[i]=(j_1,j_2)$,$j_1=1,2,j_2=1,2\cdots34$,对于省级层面,调用了省级变量,各省的人均地区生产总值 $\mathrm{GDP}_{j2[i]}$,用以估算各省之间的变异量。

对于 $j=(j_1,j_2)$ 代表的组,记该组内甜味偏好非中立($\mathrm{sweet}_i\neq0$)的样本点个数为 n_j,偏好甜($\mathrm{sweet}_i>0$)的样本点个数记为 y_j,一个自然的模型是 $y_j\sim\mathrm{Binomial}(n_j,\theta_j)$,参数 θ_j 刻画了该组人群偏好甜的概率(即 $\theta_j=\Pr(\mathrm{sweet}_i>0\mid\mathrm{sweet}_i\neq0)$)。

在变量的选择上,首先,在多层次回归模型中,加入随机效应(random effects)的回归项 α_S,即随着性别和省份变动的截距项 $\alpha_{1[j_1]}$,$\alpha_{2[j_2]}$。

为了刻画在"省级"以上的大范围地理区域的影响,我们再引进第三个分类变量:各省所属的地区(华东、华南、华北、华中、东北、西北、西南、港澳台),由此加入截距项 $\alpha_{3[j_2]}$。

其次,考虑各个变量间的交叉作用,将这些变量相互交叉,形成更多的交叉项 α_S,S 是 $J=\{1,2,3\}$ 子集,但是显然地区不应该和省份交叉,故 $S\in\{\{1\},\{2\},\{3\},\{12\},\{13\}\}$。

最后,再加入线性回归部分 GDP_{j2},并且设定其回归系数随着性别而变动。

综合以上,最终的模型是:

$$y_j\sim\mathrm{Binomial}(n_j,\theta_j),$$

$$\theta_j=\mathrm{logit}^{-1}(\alpha_0+\alpha_{1[j_1]}+\alpha_{2[j_2]}+\alpha_{3[j_2]}+\alpha_{12[j]}+\alpha_{13[j]}+\beta_{[j_1]}\mathrm{GDP}_{j2}),$$

对于以上参数,分别给定弱信息先验:

$$\alpha_s,\beta\sim N(0,\sigma^2),\ S\in\{\{1\},\{2\},\{12\},\{13\}\},$$

对于 σ^2,给定一个超先验(hyperprior):

$$\sigma^2\sim\chi^{-2}(\upsilon,{\sigma_0}^2),$$

计算此模型需要估计的回归系数共有 130 个(β:2,α_1:2,α_2:34,α_3:8,α_{12}:2×34,α_{13}:2×8)。如果使用传统的 MCMC 方法和软件,例如 BUGS,收敛速度太慢,改用 Stan(Stan Development Team,2014)建模语言进行拟合,此软件使用 H-MCMC 方法(Hoffman 和 Gelman,2013:1593 – 1623),在计算性能上更优。经过 10 条链,各 5 000 次迭代之后,所有回归系数的 Gelman-Rubin $\hat{R}\approx1$(Gelamn 等,2003),可见收敛性很好。

① 我们缺乏各省各手机平台男女用户比例可靠的总体数据,因此在以下分层分析中,不纳入手机平台这个变量。

以上的计算,也可以轻易地透过 R 软件的 lme4 包(Bates 等,2014: 1-7)实现。具体的 R 程式码如下:

MRP. fit←glmer(y ~ gender* gdp + (1 + gender | region) + (1 + gender | province) + (1 | gender. region) + (1 | gender. province) + (1|id), family = binomial(link = "logit"))

经过比较,两种方式获得的估计值 $\hat{\theta}_j$ 在统计检验上并无明显差异。

最后,使用全国第六次人口普查数据,获得各个分层的人口比例,以此为权重,对拟合值 $\hat{\theta}_j$ 进行事后加权,如以下式子:

$$\hat{\theta}_S = \sum_{j \in J_s} N_j \theta_j / \sum_{j \in J_s} N_j$$

得到各个分层民众对于粽子甜咸口味偏好的概率估计值 $\hat{\theta}_S$,而 $\hat{\theta}_S$ 的均值即是全国民众对于粽子甜咸口味偏好的概率估计值。

图 3 的第二列两幅地图展示了分层结果和全国平均值。灰色代表加权得到的全国平均值 54%,越深的蓝色/棕色,则代表了该组拟合值 $\hat{\theta}_S$ 比全国均值高/低的越多。对比原始数据和拟合值,我们看见经过 MRP 方法调整之后的拟合值在相邻组间更为平滑,也就是减少了小样本极端值带来的影响。图 4 展示了部分组在调整前后 $\hat{\theta}_j$ 的差异,可以看出对于样本量小的分层(例如宁夏、青海、西藏等地区),再经过 MRP 方法的调整后,估计值会靠近总体的均值。而对于样本量大的分层(广东、山东、河南等地区),因为包含了足够多的信息,调整前后几乎差异不大。从实际意义来说,这样的结果也更加合理。透过分层,我们得到了全国各地粽子咸甜的精确分布和男女差异,比较图 3 可以看出,北方各省男性民众的甜度指数明显不如北方各省女性民众的甜度指数,反映了一般来说女性较男性嗜吃甜食的传统印象。

总的来说,最后得出的粽子口味地理分布地图,远比所谓"南咸北甜"的通俗观点来得精细。仅就均值而言,在我们原始数据中,人们是稍微偏好咸粽的(甜味指数 0.49),这与先前其他的网络调查所得的结果是吻合的(淘宝销售数据: 39% 粽子销售为甜粽,新浪微博投票数据: 45% 的中国人偏好甜粽)。但是由于网络用户的空间布局并不均匀,使得这样的结果有偏,经过 MRP 调整后,中国民众反而稍微偏好甜粽(甜味指数 0.54)。

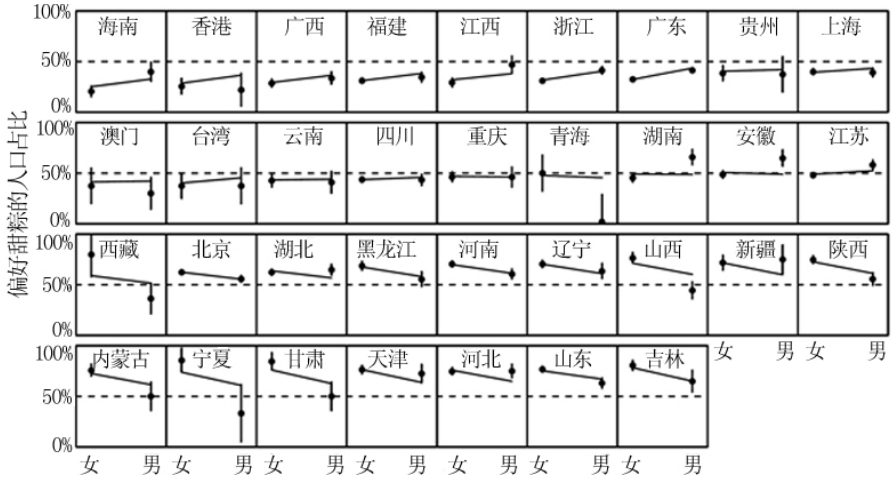

图 4 各个分层原始数据 θ_j 与经过多层次回归调整后的 $\hat{\theta}_j$ 的比较

注: 图中实心点代表原始数据 θ_j,垂直线代表 ±1 个标准误,越长的垂直线反映了该组样本量越小,实线代表经过多层次回归调整后的 $\hat{\theta}_j$,地区依照 $\hat{\theta}_j$ 的均值由左上至右下进行排序: 海南民众对粽子的口味偏好最甜,而吉林民众对粽子的口味偏好最咸。

六、结　论

本文使用 MRP 方法,结合 Rweibo 包自动抓取微博文本的功能,实现利用网络大数据进行舆情研究,并从不具代表性的互联网抽样数据,调整加权得到更加精确的分组估计值,以及更合理的全国均值。我们认为,稍加替换微博搜索的关键词,本文所使用的 MRP 方法,便可以推广到其他议题,研究者可以对任何感兴趣的舆情做出更精细、更合理的估计。本文提供的方法具有一定的普适性,是可以进一步推广的。在本文使用的例子中,仅在分层时考虑了地区和性别这两类变量,这是因为新浪微博目前仅开放这两种变量的接口,将来一旦我们可从获得更多的用户个人信息,或者有其他网络平台可以提供类似新浪微博,但是用户信息含量更多的数据,MRP 方法可以毫不费力增加分层的变量,让研究者可以更为省时、省力地获得更为精细、信息更为丰富的估计值。

参考文献:

[1] 金勇进、张喆,2014,《抽样调查中的权数问题研究》,《统计研究》第 9 期。

[2] 邱东,2014,《大数据时代对统计学的挑战》,《统计研究》第 1 期。

[3] 朱建平、章贵军、刘晓葳,2014,《大数据时代下数据分析理念的辨析》,《统计研究》第 2 期。

[4] Bates, Douglas, Martin Maechler, Ben Bolker, et al, 2014, lme4: Linear Mixed-Effects Models Using Eigen and S4, R package Version 1.

[5] Gelman, Andrew, John B. Carlin, Hall S. Stern, et al, 2003, *Bayesian Data Analysis*, 2nd edition, London: CRC Press.

[6] Ghitza, Yair, and Andrew Gelman, 2013, Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups, *American Journal of Political Science*, 57(3).

[7] Hoffman, Matthew D., and Andrew Gelman, 2013, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research*, 15.

[8] Li, Jian, and Yibo Chen, 2014, Rweibo: An Interface to the Weibo Open Platform, R package version 0.2-9/r49.

[9] R Core Team, 2014, R: A Language and Environment for Statistical Computing, http://www.R-project.org/.

[10] Stan Development Team, 2014, Stan: A C++ Library for Probability and Sampling, Version 2.4, http://mc-stan.org.

[11] Wang, Wei, David Rothschild, Sharad Goel, et al, 2014, Forecasting Elections with Non-Representative Polls, *International Journal of Forecasting*, Forthcoming.

(责任编辑: 匡　云)