台湾大选与选举预测

徐轶青

台湾大选结果刚刚公布。初步开票结果，蔡英文以 56%当选台湾地区领导人，国民党候选人朱立伦惨败，得票率为 31%。因为台湾大选的关系，不少朋友开始对选举预测感兴趣，并在社交媒体上展开了热烈的讨论，也产生了一些误解。

我虽然自己不做这方面的研究，但是了解其中的基本原理，可以抛砖引玉地做一些解释。

从某种程度上说，选举预测（以及其他与选举相关的数据和分析工作）可以说是美国政治学的生命线。因此，在斯坦福学习期间，我专门旁听了 Simon Jackman 和 Douglass Rivers 两位教授的课程。前者开发和完善了目前最常用的选举预测模型，许多大报用的就是 Simon 的模型；后者更是资深的业界大拿，领导着 YouGov 的选举预测团队。我从他们那里学到很多。

选举预测，说难很难，说简单也很简单。

先说它为什么简单。因为预测的目标是一个平均数。比如说，一个选民要么投蔡英文（记为 1），要么不投她（记为 0）。蔡英文的得票率，就是上千万台湾选民投票结果的平均数。

因为统计里有“大数定理”和“中央极限定理”，在理想情况下，这个平均数是非常容易预测的。我们可以抽样的办法，从上千万人里随机里抽 500 个人，用他们对蔡英文的支持率来推测蔡英文的支持率（得票率）。

这有些令人难以置信。我们怎么可能通过 500 个人来判断上千万人行为的结果呢？

是难以置信。2008 年（甚至到 2012 年时也是如此），美国的一些共和党人也有类似的困惑：怎么可能通过基于 1000 人的民调推断近 3 亿人行为的结果呢？在投票还没有结束、一个票箱都还没开的时候，电视台怎么就能判断某州是民主党的呢？当共和党被打得满地找牙的时候，他们就不得不相信科学的力量了。

一些选举的制度设计，可能使选举预测变得更简单。比如美国的选票人票制度（决定最后谁当总统）：当一党候选人得到该州的多数票时，该州所有的选举人票都归其所有。在过去历次总统选举中，民主党总统候选人在麻州的得票率总是远高于 50%。这就是为什么，不用开票、不用反复做民调就知道，麻州的 10 或 11 张选举人票铁定是归民主党的。不用预测。正因如此，两党的候选人也不会在麻州花钱花时间。

我们再来讨论，选举预测难在哪里。如同所有预测问题，选举预测的难点有两个，一是信息不足，二是样本偏误。

如果一个选举在历史上已经发生过很多次了，那么类似的选举就容易预测。原因如下。选举预测，一般利用两部分信息。一是人口的基本面信息，如年龄、性别、种族、职业、宗教信仰、受教育程度、收入状况、甚至包括性取向，当然还包括党派属性和过去的投票经历。二是来自民调（抽样调查）的信息。

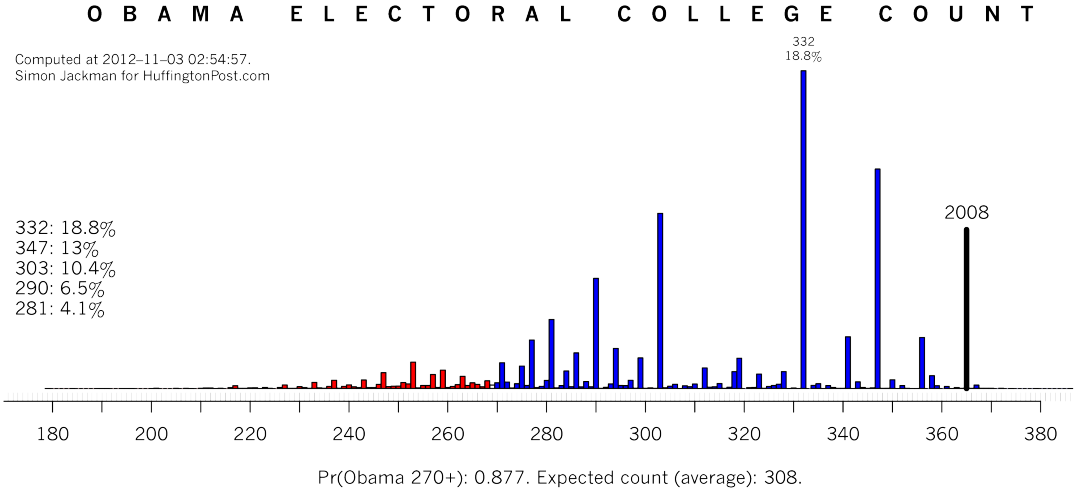
比方说，一个从哥伦比亚大学毕业、住在纽约的黑人白领女性，十有八九是民主党的支持者。一个高中毕业、住在德克萨斯农村的白人男性福音派信徒，十有八九是共和党的支持者。为什么？深一点的答案与两党的意识形态和政策定位有关；浅一点的答案：过去的大选数据一再验证了这种相关性。

如果一种选举是第一次发生，比如 2014 年的苏格兰独立公投，我们不太清楚基本面信息与投票之间的相关性，选举预测就变得相当困难。只能够更大程度地依赖民调。

人口基本面的情况变化慢，但是对预测的解释力更强、更稳定。相反，短期内，民意波动的幅度很大。研究者发现，在离选举较远时，真金白银的预测市场（其实就是以选举结果定胜负的赌场）的预测能力要远胜过民调。但是，越接近选举，民调就变得越准、越重要。因此，我们可以把基本面信息和民调分别理解为“趋势”和“波动”。两种信息对选举预测都很重要。

目前比较流行的（也被证明确实有效的）模型，是通过贝叶斯法把两类信息结合起来，然后随着一个个民调出炉，不断更新对最终结果的预测。当然，需要一些经验来设定模型的参数，比如新的信息相对已经积累的信息有多重要。

像美国大选这样已经积累了许多数据的选举，对于选举人票的预测已经可以做到八九不离十。2012 年，美国有三支团队独立地预测对了所有州选举人票的归属，其中就有 Simon Jackman（下图）和在坊间更有名的 Nate Silver。我已经粗略解释了，这并不是变魔术，但是确实需要非常细致、认真的数据搜集和分析工作。



第二种使选举预测变得比较困难的因素是样本偏误。在大部分国家和地区，参加投票都不是强制性的。那么，对候选人来说，鼓动他们的铁杆支持者去投票站投票（台湾人叫“催票”）就变得特别重要。对预测者来说，就需要先预测谁会出来投票，再预测他们会投给谁——或者等价地，预测很可能出来投票的人（likely voters）会把票投给谁。

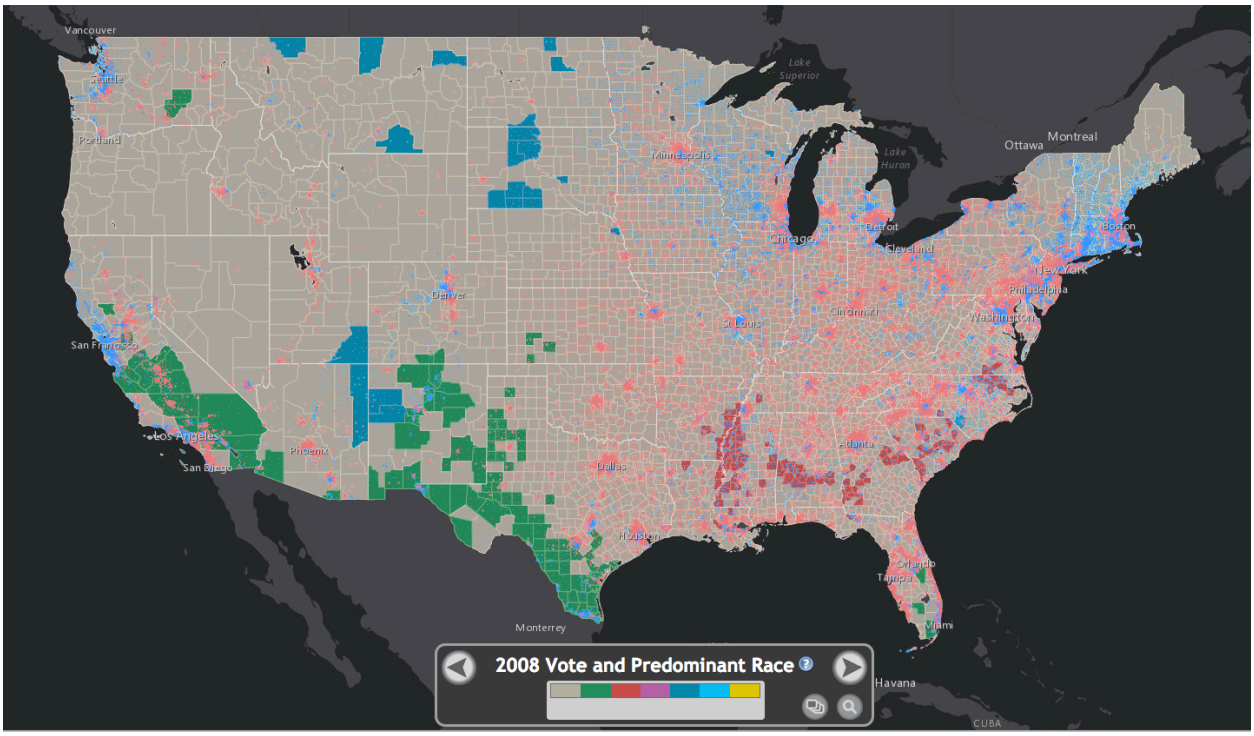
目前，对于谁会出来投票，并没有特别好的技术来预测。投票率（turnout）受到天气、交通、心情的影响，也受到选民认为他或她的那票有多重要的影响。如果选民认为自己支持的候选人铁定会赢或者铁定会输，就很可能不出来投票了。所以，在做民调时，一般会直接问受访者是否考虑去投票，但这并不一定准。投票率可能会受到一些突发事件的影响，如 2004 年的“两颗子弹”。

顺便提一句，过去做民调，一般用座机。现在大部分人改用手机了，又增加了一重抽样的难度。

另外，现在像 facebook、twitter 这样的社交媒体也越来越多地参与选举预测，或给选举预测团队提供信息，因为他们掌握着很多类似民调的波动信息。目前，预测者之所以对社交网络上的信息用得比较少，有两个原因。一是这些站点的样本偏误可能很严重。二是，做严格科学抽样的民调其实并不贵。

还剩下一个技术问题。从直觉上我们就知道，预测某选区立委选举的结果，要比预测大选结果困难不少。原因不仅是单个选区噪音更大，还因为大部分民调都是针对大选而非个别选区抽样的。统计学和政治学家们开发出一种他们称为 MRP 的贝叶斯估计法来解决这个问题，其基本思路也是借助基本面信息来推测某特定选区选民的投票倾向。

小结一下。第一、选举预测没有什么神秘的，就是预测一个平均数。第二、一般利用两部分信息：基本面管大势，民调管波动，然后把两个结合起来。目前，选举预测技术已经日臻完善（差 5%就是两个标准误以外了），关键是要搜集更多的人口基本面信息（如下图的斯坦福 atlas 数据库，底色表示当地多数种族）和做更多民调、或类似民调的东西。



如果有民调的信息放着不用，就是直接丢弃重要的波动信息，会影响预测的质量。而且，直观上仿真并不能解决信息缺失的问题。我认为，对选举预测技术的创新，需要在公开、透明的环境下评估，也能知道新方法相对于已有的方法是否有所改进。

为什么没有很多大陆学者参与台湾的选举预测？我想有两个原因。一是台湾同行已经做得很不错了（例如我的一位台湾朋友领导的“政治大学选举研究中心”）。二是大家可能认为，从学术和政策的角度，做这方面研究的收益不大。不过我觉得，选举预测是政治学里较接近科学的部分，希望更多人能参与进来。