

量刑自由裁量权的边界：集体经验、个体决策与偏差识别

吴雨豪^{*}

内容提要：在量刑规范化改革中，如何在赋予法官自由裁量权的前提下，将“案件由谁审理”这一因素所导致的量刑偏差控制在合理限度内，是理论和实务要解决的问题。一种可行的方案是，将法官个体的刑罚裁量与法官量刑集体经验进行对比，对靠近集体经验量刑的法官的自由裁量权采取肯定和尊重的态度，而对量刑显著偏离集体经验的法官的量刑决策进行识别并纠正其偏差。基于北京地区五类案件近5万份刑事判决书，对盗窃、诈骗、抢劫、故意伤害和交通肇事案件的刑罚裁量展开实证研究后发现：“案件由谁审理”这一变量能够解释5%—11%的量刑差异；大多数法官的量刑结果在集体经验附近浮动，只有少部分法官的量刑结果显著偏离集体经验；更有少部分法官在多个罪名的刑罚裁量中均出现量刑偏差。量刑结果显著偏离集体经验的法官应成为自由裁量权规制中重点关注的对象，尤其是在多个罪名的刑罚裁量中均出现偏差的法官。量刑自由裁量权的规制方案，彰显了大数据与算法在量刑规范化改革中的应用前景，在引入“算法的程序控制”理念的前提下，具有推广价值。

关键词：量刑规范化 自由裁量权 集体经验 偏差识别

一、问题的提出

在罪刑相适应原则的制约下，法官群体的量刑决策应当具有稳定性。犯罪严重程度、责任刑情节轻重相近的被告人，即使由不同法官进行裁判，其量刑结果应当相近。这不但是司法公正的具体表现，也符合人类朴素的正义观念。2009年，最高人民法院发布《人民法院第三个五年改革纲要（2009—2013）》，提出规范自由裁量权，研究制定量刑指导意见，由此开启了量刑规范化改革的序幕。此后，我国的量刑指导意见历经多次修订，成为影响基层法官量刑行为最重要的规范性文件之一。实证研究也显示，在量刑指导意见颁布后，基层法院刑罚裁量的明确性和稳定性均显著提高。^{〔1〕}

^{*} 北京大学国际法学院助理教授。

〔1〕 参见白建军：《基于法官集体经验的量刑预测研究》，《法学研究》2016年第6期，第153页；王越：《量刑规范性水平的实证检验：以故意伤害罪为例的分析》，《法学家》2020年第6期，第77页以下。

然而,不可否认的是,即使规范性文件对量刑步骤和方法作了精细化规定,在司法实践中,量刑仍然是法官基于自由裁量权的个体决策行为。一方面,虽然量刑指导意见对所要考虑的量刑情节和裁量幅度作了列举式规定,但具体到每一个量刑情节,量刑指导意见仍然赋予了法官相当的自由裁量空间。^{〔2〕}当多个这样的情节叠加时,自由裁量权所引发的量刑结果差异就相当可观。另一方面,量刑决策是规范判断和价值判断的有机统一,^{〔3〕}法官作为社会公众的一员,在裁量刑罚时,不可避免地会受到自己的生活常识、司法潜见乃至刑事政策价值取向的影响。^{〔4〕}这些非法律因素很可能隐秘地存在于量刑决策的过程中,最终对量刑结果产生显著影响。

由于上述原因的存在,法官自由裁量权所导致的量刑差别广泛存在于司法实践中。但是,如果量刑差别超过了合理限度,势必会对司法公信力产生冲击。首先,根据责任主义的原则,被告人承受的刑罚应当与其行为的法益侵害程度相适应。^{〔5〕}在其他因素相同的情况下,具有相似违法事实的被告人应当承担相似程度的刑事责任。如果法官个人因素对量刑的影响超过一定的幅度,“案件由谁审理”这一与犯罪事实不相关的因素就会显著影响量刑结果,导致同案不同判的结局,进而违背报应主义的观念。其次,根据特殊预防理论,刑罚处罚的目的之一是通过惩戒和教育手段降低犯罪人的再犯危险性。如果被告人意识到决定自己刑罚轻重的因素中有相当一部分并不来源于自己的犯罪事实和量刑情节,其就有可能不再信服自己所承担的刑事责任的正当性和合理性,从而导致刑罚的特殊预防功能落空。最后,根据积极的一般预防理论,刑罚的任务是“在法律共同体中证明法律秩序的牢不可破,并且由此加强人民的法律忠诚感”。^{〔6〕}如果量刑轻重过度依赖裁判者的个人因素,社会一般公众就会认为法律的适用缺乏稳定性和可预期性,这对于法忠诚的信念树立会有不利影响。

因此,一个合理的刑罚裁量体系应当在赋予法官自由裁量权的前提下,将“案件由谁审理”所导致的量刑偏差控制在合理限度内。这不但是量刑规范化的制度要求,也是刑罚功能和目的的应然之理。由此引出两个基础性问题:第一,如何识别自由裁量权所导致的量刑偏差;第二,以何种方法定义“合理限度”。

一方面,与定罪决策中事实与规范之间的涵摄关系不同,在量刑决策中,法官不但要在定性上考虑各种因素的有无,还要在定量上权衡各种因素影响量刑的程度。由于多种情节和影响幅度组合的复杂性,量刑裁量过程存在难以计数的可能性。同时,量刑的结果输出又是一个简单的数值,在遵循法定刑幅度和量刑指导意见的前提下,量刑结果妥当与否,并没有进一步的规范判断标准,从而为纯粹从法律推理角度界定量刑偏差设置了障碍。因此,不得不转换视角,从经验和事实的角度定义量刑偏差。本文提出,法官量刑的集体经验可以成为衡量量刑个体决策的重要参照系。基于量刑规范化和同案同判的理念,在控制了所有量刑情节之后,如果某一法官的量刑结果仍然与法官量刑集体经验存在显著偏离,这种偏离就可以定义为个人因素

〔2〕 例如,量刑指导意见规定,已满16周岁不满18周岁的未成年人犯罪,减少基准刑的10%到50%。如果是基准刑为10年有期徒刑的未成年人被告人,仅就该量刑情节,不同案件中的基准刑减少幅度就可以横跨1年至5年。

〔3〕 参见彭文华:《量刑的价值判断与公正量刑的途径》,《现代法学》2015年第2期,第112页。

〔4〕 See B. D. Johnson, *Judges on Trial: A Reexamination of Judicial Race and Gender Effects across Modes of Conviction*, 25 (2) Criminal Justice Policy Review 159-184 (2014); M. Schanzenbach, *Racial Disparities, Judge Characteristics, and Standards of Review in Sentencing*, 171 (1) Journal of Institutional and Theoretical Economics 27-47 (2015).

〔5〕 参见张明楷:《责任刑与预防刑》,北京大学出版社2015年版,第275页。

〔6〕 [德] 克劳斯·罗克辛:《德国刑法学总论》第1卷,王世洲译,法律出版社2005年版,第42页。

所导致的量刑偏差。

另一方面，法官的刑罚裁量不是一个纯粹机械的行为，刑罚个别化原理要求法官在量刑时必须考虑个案的情境因素，作出符合个案正义的判决。因此，即使在控制了所有可能的法定量刑情节之后，由于自由裁量权的存在，所有的量刑结果也不可能严格趋同于一个所谓的符合集体经验的特定数值。所以，在量刑决策的规制中，又需要为个体决策与集体经验的偏离赋予一定的容忍度。其体现在操作层面，就是要为上述量刑偏离设定阈值。在这个阈值内，刑罚裁量偏离集体经验会被认定为自由裁量权的合理体现，法官的个体决策应当受到肯定和尊重。若超出这个阈值，则自由裁量权的行使超越合理限度，存在被滥用的风险，应当予以识别并纠正其偏差。

综上所述，在对法官量刑自由裁量权进行规制的过程中，需要比对个体的量刑决策与群体的集体经验，运用定量研究方法计算量刑偏离，并判断其是否超越合理限度。在明晰了这一基础性问题后，本文将以裁判文书数据为样本，基于特定的统计算法展开实证研究。

二、作为参照系的量刑集体经验

（一）量刑集体经验的测量与呈现：从均值估计到量刑分布

白建军最早提出量刑集体经验的概念，并将其表述为“司法案件大样本中实际法律问题与法律实践之间实然联系的科学综合与归纳”。^{〔7〕}在具体操作上，白建军基于裁判文书数据，以法定情节为自变量、以刑期为因变量，对量刑大样本进行回归分析，得到刑期与法定情节之间关系的统计模型。^{〔8〕}这一统计模型本身，构成了法官集体经验的具体呈现方式。以交通肇事案件为例，如果统计模型的计算结果显示，“有逃逸、负事故全部责任、致死一人等三个情节”的预测值为有期徒刑 1036 天，这一结果就是量刑实践中法官集体经验的具体呈现。^{〔9〕}

本文遵循上述思路，认为从司法裁判大样本中总结出的量刑统计模型能够测量和呈现量刑集体经验。然而，一个无法忽视的事实是，如果仅简单地以多元线性回归模型作为集体经验的载体，研究者不可避免会忽视蕴藏于个体决策中的量刑行为的复杂性。这是因为，在统计学意义上，线性回归所呈现的结果只是对集体经验的均值估计。^{〔10〕}例如，如果量刑模型发现，累犯情节会导致 20% 的刑罚加重效应，那么，其真实意义是，从平均意义上说，被告人是累犯这一事实会使其刑罚加重 20%。但事实上，一些被告人的累犯情节所对应的刑罚加重幅度可能低于这个值，也可能高于这个值。同样，量刑预测模型输出预测结果，也只是特定情形下量刑结果的均值估计。例如，在前述交通肇事案件“有逃逸、负事故全部责任、致死一人等三个情节”的情形下，虽然模型预测的有期徒刑均值是 1036 天，但白建军同样发现，在实际的司法裁判中，部分裁判结果远低于这个模型预测值（如 510 天），另一些裁判结果则显著高于这个预测值（如 1530 天）。^{〔11〕}这两个极端结果显然无法为均值估计所涵盖。

〔7〕 前引〔1〕，白建军文，第 152 页。

〔8〕 同上文，第 148 页。

〔9〕 近期学界出现了多篇运用类似研究方法的文章。这些研究均在采集大样本裁判文书的基础上，通过模型总结出量刑情节与量刑结果的统计关系，并将量刑模型中体现的统计关系作为我国量刑实践经验的归纳和总结。参见文姬：《基准刑调节方法实证研究》，《法学》2020 年第 2 期；前引〔1〕，王越文；彭雅丽：《量刑指导意见的司法实践与重构——以盗窃罪为切入点》，《法学研究》2021 年第 4 期。

〔10〕 参见谢宇：《回归分析》，社会科学文献出版社 2013 年版，第 50 页。

〔11〕 参见前引〔1〕，白建军文，第 149 页。

更为重要的是,从量刑模型中获得的均值估计,无法实现本文期望达到的“通过集体经验识别个体决策偏差”的研究目标。例如,量刑模型呈现,在控制了所有量刑情节后,法官量刑实践中的集体经验将某一罪名的基准刑设定为3年有期徒刑,即使发现一位特定的法官将量刑基准刑的均值设定为4年有期徒刑,在法官量刑结果聚集的情况下,该法官个体决策的偏离程度(高于集体经验均值1年的量刑偏离)也可能已经超越了合理限度;而在量刑结果分散时,1年的量刑偏离也可能仍然在自由裁量权的合理限度内。

因此,本文认为,一个完整的量刑集体经验的建构,应当同时包括对量刑实践中均值的测量和对个体决策离散程度的呈现。由此,量刑集体经验的测度分为两个部分:(1)在量刑大样本下,控制了所有量刑情节之后,通过一般化量刑模型,测量出法官集体的量刑均值。这一量刑均值将是量刑集体经验的基准。(2)在量刑模型中加入法官个体决策的因素,计算出每位法官相对于集体经验均值的偏离值,并由此得到关于所有法官个体决策的量刑分布。^[12]例如,基于某一罪名的量刑大样本,在通过量刑模型控制了所有可能的量刑情节之后,得出“所有法官量刑结果的均值为3年有期徒刑”的结论。“3年有期徒刑”就是该罪量刑集体经验的均值。如果恰好有一位法官的量刑结果等于这个均值,这位法官的个体决策与集体经验不存在任何偏离,该法官就会被我们认定为“基准法官”。而对于大多数法官,其量刑决策与集体经验均值之间都会存在一定程度的偏离,这种偏离就会被认定为相对于集体经验的量刑偏离。例如,量刑均值为4年有期徒刑的法官决策,相对于集体经验的量刑偏离就是1年。同时,由于量刑分布汇集了所有法官的量刑决策,我们就能够从该分布中获取有关该法官量刑偏离严重程度的信息。比如,如果我们发现该法官的量刑结果虽然与集体经验均值相差1年,但其量刑决策仍然处于量刑分布中较为中心的位置,就意味着该法官自由裁量权的行使很可能并未超越合理限度(见图1)。反之,如果我们发现该法官的量刑结果显著靠近量刑分布的两端,其自由裁量权的行使就很可能已经超越了合理边界。

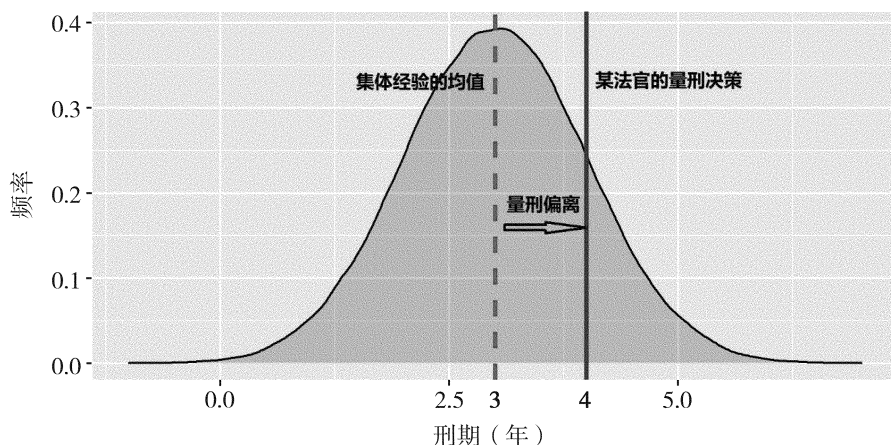


图1 法官集体的量刑分布

(二) 通过集体经验规制个体决策:蕴藏于实然中的应然

不难看出,当量刑集体经验能够同时以“集体经验均值”和“个体决策分布”两个方面

[12] 本文会在统计方法部分对这两个步骤的技术实现方式作详细论述。

呈现时，本文倡导的“通过集体经验规制个体决策”就有了可行性。因为，一方面，我们可以将每位法官的量刑结果与法官群体的集体经验均值进行对比，计算出个体决策相对于集体经验的量刑偏离值。另一方面，我们能够从所有法官个体决策的分布中，获得这种偏离是否严重的信息，并最终将之作为量刑自由裁量权是否超过合理限度的判断依据：当某一个体决策处在集体经验分布中较中心的位置时，即使其与集体经验均值存在一定的偏离，也应被认为属于合理行使自由裁量权；反之，如果个体决策处在集体经验分布中的极端位置，则自由裁量权存在被滥用的风险。在这一宏观思路下，关于量刑集体经验这一参照系的建构，仍有以下三点需要专门说明。

首先，量刑集体经验的建构，需要建立在尽可能排除个案差异的基础上。由于一系列法定和酌定量刑情节对量刑结果具有决定性作用，即使发现不同量刑情节组合的案件具有不同的量刑结果，也无法将之归结为不同量刑主体的决策差异。因此，无论集体经验均值，还是法官量刑决策分布，都需要排除量刑情节的客观影响，从而使不同量刑主体能够在相同的标准下公平比较。幸运的是，统计学工具的发展能够帮助我们实现这一目标，使我们能够在控制了所有可能的量刑情节之后，完成量刑集体经验的建构。

其次，量刑集体经验与个体决策之间，存在着一种整体与个体的辩证关系。一方面，整体的方向是由所有个体的合力方向牵引，量刑集体经验的建构建立在对大量具体量刑决策进行综合与归纳的基础之上。因此，集体经验本身并不先于个体决策而存在，没有法官群体量刑实践的聚沙成塔，就不存在所谓的量刑集体经验。另一方面，对于特定个体的量刑决策而言，其在整体中的位置又具有重要的评价意义。因为，在大数据的语境下，法官的量刑实践必然存在着向集体经验中央聚集的趋势（后文的实证研究也证实了这一形态的客观存在）。这一中间聚集的规律，并不会因为少量的量刑决策而改变。因此，在评价某个具体量刑决策时，衡量其对于量刑集体经验中央聚集趋势的偏离程度，具有重要的现实意义。

事实上，这种通过衡量个体在整体中的相对位置，从而得出评价结论的思维方法，在社会科学领域已经被广泛运用。例如，在社会学领域，有学者将城镇居民年人均收入的 25% 分位点与 95% 分位点作为定义我国城镇居民中等收入群体的下限和上限。^[13] 这意味着，一方面，我国城镇居民收入是由大量个体收入情况汇总而成的。另一方面，这一收入分布对于衡量个体的收入水平具有重要意义，当某一个体的收入水平处于整体收入分布的极端位置时，例如低于收入的 25% 分位点，该个体就会被定义为低收入人群。在刑事司法领域，也已经有学者通过这种内部基准法（internal benchmarking）来衡量执法人员的自由裁量权。例如，里奇韦和麦克唐纳以美国辛辛那提市的警察执法为背景，衡量个体执法行为在整体执法状况中的相对位置，并以此评价执法是否公正。^[14] 这与本文“通过集体经验规制个体决策”的设计思路不谋而合。

最后，一种可能存在的对“通过集体经验规制个体决策”思路的顾虑是，作为参照系的集体经验同样是对量刑经验的实然总结，而量刑自由裁量权规制要回答的却是“刑罚裁量应当怎么样”的问题。换言之，以实然的集体经验为判断标准，为何能得出一个应然的结论？

一方面，无论是罪刑相适应原则，还是量刑指导意见中体现的“对于同一地区同一时期案情相似的案件，所判处的刑罚应当基本均衡”的理念，都要求量刑情节组合相似的被告人，

[13] 参见李培林、朱迪：《努力形成橄榄型分配格局——基于 2006—2013 年中国社会状况调查数据的分析》，《中国社会科学》2015 年第 1 期，第 51 页。

[14] See G. Ridgeway & J. M. MacDonald, *A Method for Internal Benchmarking of Criminal Justice System Performance*, 60 (1) *Crime & Delinquency* 145–162 (2014).

即使由不同法官进行裁判,也应该得到相似的量刑结果。因此,在认为刑罚裁量的稳定性和法的确定性是一个应然的价值诉求的前提下,在控制了量刑情节之后,类似案件中不同法官个体的刑罚裁量结果应当相近。在具体形态上,个体决策应当呈现出向集体经验中央聚拢的趋势。反之,如果个体决策之间的差异过大,或者某一决策过于偏离集体经验均值,就可能预示着刑罚裁量违背上述应然价值。

另一方面,虽然量刑指导意见对量刑基准和量刑情节的作用幅度作了具体规定,但量刑指导意见也提出“量刑要客观、全面把握不同时期不同地区的经济社会发展和治安形势的变化”。由此可见,在一个应然的刑罚裁量体系中,量刑决策不但要符合实体量刑规范规定的幅度,还要具有回应社会现实之审时度势、因地制宜的特性。事实上,既有的实证研究已经发现,如果仅以量刑是否符合具体规范作为评价标准,法官的量刑决策大多会在规范规定的幅度之内。^[15]因此,评价量刑是否正当,就聚焦到“量刑是否符合社会现实”的考量。

本文认为,正是在这个意义上,从法官量刑实践中总结出的集体经验具有重要价值。法官作为刑罚裁量的主体,其会在日常的司法实践中直观地感受到社会经济变化、犯罪治理状况、民众态度的变更等一系列可能影响刑罚裁量的社会因素。由于司法活动是一个能动的过程,法官在具体案件量刑过程中会对这些因素作出价值判断,并由此根据经验对量刑决策作出调整。^[16]当多个这样的调整形成合力时,个体的选择就汇聚成了集体智慧,从而反过来成为量刑决策是否合理的重要参照对象。例如,量刑指导意见将累犯加重的幅度设定为10%—40%。设想随着社会治安状况的改善、一般预防必要性的降低,累犯问题不再突出,法官群体出于刑罚经济性的考虑,会更倾向于靠近下限10%进行量刑,而此时如果有一位法官仍将加重的幅度设定为40%,那么,这一个体决策虽然没有违背实体规范,却不符合法官的集体经验、也不利于这种情形下刑罚目的的实现,因此同样要受到规制和纠正。这就是所谓“用经验去校准经验”的意涵。^[17]

三、研究假设与检验方法

(一) 检验逻辑

在明晰了“通过集体经验规制个体决策”的基本思路后,对量刑自由裁量权边界的检验分为宏观和微观两个层次。宏观上,本文将检验,在一个刑事司法体系内部,法官的自由裁量权总体上对量刑结果会产生何种程度的影响。微观上,本文将定位于每一位具体的法官,将每一位法官的量刑决策放在法官量刑的集体经验中进行考察,并检验特定法官的量刑自由裁量权是否超越了合理边界。

这两个层面的问题看似关注的对象不同,实则前后递进。宏观层面的检验立足于发现问题。如果发现法官个人因素对量刑的影响具有普遍性,我们才要反思如何设定这一影响的合理边界。微观层面的检验立足于解决问题。因为只有发现特定法官的量刑偏差,我们才能创建一套规制量刑自由裁量权的识别和纠偏系统,进而从整体上减少法官个人因素对量刑公正产生的

[15] 参见前引[1],王越文,第77页以下;前引[9],彭雅丽文,第177页以下。

[16] 参见周光权:《量刑规范化:可行性与难题》,《法律适用》2004年第4期,第61页。

[17] 参见前引[1],白建军文,第154页。

冲击。依循上述逻辑，本文将依次检验三个假设：

假设1：宏观上，如果法官个人因素是影响量刑的重要因素，那么，“案件由谁审理”这一要素将能解释个案之间一部分的量刑差异。

在既有的实证研究中，学者广泛使用量刑预测模型来刻画量刑情节与量刑结果之间的关系。但是，无论研究者如何穷尽各种量刑情节，从量刑模型中推导出的预测值与实际刑期，在个案中依然会存在一定的差距。这种残差可能是各种复杂结果的结果，包括未写入判决书的法律因素与其他非法律因素。^{〔18〕}在这些因素之外，还有一个可能的重要原因——不同法官的量刑裁量模式不同。因此，如果认为法官自由裁量权是影响量刑的重要因素，那么，在量刑情节之外，“案件由谁审理”将能解释一部分的量刑差异。

假设2：微观上，某一法官的量刑越是靠近法官集体量刑分布的某一极端，就意味着该法官个人因素对量刑结果的影响越有可能超过合理边界。

正如前文所示，在同案同判和量刑稳定性的价值指引下，在控制了相关量刑情节之后，法官的量刑决策应当呈现出向集体经验均值聚拢的趋势。而对于每一位特定的法官，我们允许其在不过分偏离集体经验的前提下行使自由裁量权。但是，如果其量刑决策过于偏离集体经验而位于法官集体量刑分布的某一极端位置，其自由裁量权就可能超越了合理边界，因而有纠偏的必要。

假设3：微观上，如果法官在不同罪名的刑罚裁量中均偏离集体经验，则对于该法官个人因素对量刑的影响更有纠偏的必要。

法官个人因素对量刑结果可能存在多种影响机制。经验表明，如果某一法官仅在某个罪名上存在显著量刑偏离，在其他罪名上则趋近集体经验，那么，我们更倾向于认为该法官只是在某一罪名的社会危害性认识或者特定量刑情节的适用上有不同的理解。然而，如果发现某一法官在所有罪名的量刑裁量上均显著偏离集体经验，那么，我们更倾向于认为该法官很可能在量刑步骤和方法、刑罚裁量的价值判断乃至刑事政策的宏观理解上存在偏差。相对于前者，后一种情形更有纠偏的必要。

值得说明的是，这三个假设虽然均来自“通过集体经验规制个体决策”的理论构想，但三者均要经过特定的实证检验后，才能转化为最终的结论。其中，假设1检验量刑大样本的总体状况，即只有数据和实证结果展示了“案件由谁审理”能够解释相当程度的量刑差异时，才能在证成假设1的基础上得出“法官个人因素是影响量刑的重要因素”的结论。假设2、假设3检验每一位法官的具体量刑决策，即只有实证结果显示某法官的量刑结果显著偏离集体经验时，才能得出“某法官的量刑结果超越合理边界”的评价结论，否则，针对特定法官量刑自由裁量权的规制也无从提起。^{〔19〕}

（二）样本介绍与变量设置

1. 样本构成

在通过集体经验规制个体决策的操作过程中，一个前置性的问题是，集体经验的范围应当如何设定。换言之，应当以什么样的量刑样本来评价具体的量刑决策。

一方面，我国幅员辽阔，各地社会经济发展水平不一；同时，在全国性的量刑指导意见之

〔18〕 参见前引〔1〕，白建军文，第147页。

〔19〕 出于对真实案件中法官隐私的保护，本文在实证结果部分并不会就“每位法官的量刑结果是否超越合理边界”的一级结论进行展示。但是，这并不代表本文没有对假设2、假设3作出实证检验。事实上，本文实证结果中“量刑偏差识别数据库的总体情况”就是在实证检验每位法官量刑决策的基础上进行的汇总。

外,各地又颁布了不同的量刑实施细则以指导当地的量刑实践。这些因素共同作用的结果是,我国法官的量刑决策很可能呈现出明显的地区差异。因此,如果将全国的量刑样本作为识别量刑偏差的参照系,一个可能的结果是,某一法官群体的量刑偏差并不是自由裁量权所导致的,而是特定司法环境或者地区性量刑规范所作用的结果。由于本文的研究目标是通过个体决策与集体经验的比对,发现量刑自由裁量权的合理边界,所以,需要划定一个相对同质化的法官集体作为识别量刑偏差的参照系。所谓同质化,是指该样本中法官集体所在的司法环境相似,处理的刑事案件类型相近,适用的量刑规范相同。只有满足这一前提,通过量刑差异识别法官自由裁量权影响程度的研究目标才能实现。因此,不同于大多数以全国范围内案件作为数据库的实证研究,^[20] 本文将以某一特定地区、同一层级法院法官审理的刑事案件为样本来源,以满足样本同质性的要求。

另一方面,既有的量刑实证研究多聚焦于某一个罪,探究法官集体在该罪中所呈现的行为模式。但是,在量刑偏差的识别中,发现某一法官在某一个罪中出现量刑偏差,其意义十分有限。一种可能的解释是,这种量刑偏差只是因为不同法官对某一个罪特定量刑情节的适用有不同理解。相反,在实践中,那些最需要引起关注的量刑决策群体,恰恰是那些面对不同类型刑事案件,其量刑始终畸重或者畸轻的法官。如同前文假设3所述,这类法官在价值取向和刑事政策选择上存在一定偏差,因而更有纠偏的必要。因此,为了构建跨案件类型的量刑偏差识别系统,本文的样本需要涵盖不同罪名的刑事案件,以此检验法官个体自由裁量权所导致的量刑偏差是否稳定地出现在不同案件类型中。

基于上述理由,本文的原始数据包括中国裁判文书网公布的所有北京市基层法院2014—2020年审理的五类刑事案件(盗窃、诈骗、故意伤害、抢劫和交通肇事)的裁判文书。第一,样本中所有的近5万个刑事案件都在北京地区审理,我们有理由期待,审理案件的法官所处的司法环境、所适用的量刑规范基本相同。第二,样本中所有的案件都来自基层法院,这将那些特别重大的刑事案件排除在外,从而增加了案件的同质性。第三,在罪名的选择上,五个罪名均属最常见的犯罪类型。这些罪名既涵盖财产犯罪、也涵盖人身犯罪,既包括故意犯罪、也包括过失犯罪,法官在这些常见犯罪中的量刑决策,可以较完整地体现其不同类型案件中的量刑偏好。

笔者通过特定的计算机软件对这些裁判文书进行了分析和读取,剔除了部分不披露具体内容的裁判文书。对于一个案件中包括数名被告人的情况,每一个被告人被单独提取作为一个独立的观测值。^[21] 最后,为了保证量刑结果衡量的同一性,本文将样本限定在犯上述罪之一并被判处有期徒刑的被告人。经过上述处理和筛选后,本文样本构成如表1所示。

表1 案件类型与主审法官分布

	盗窃罪	诈骗罪	故意伤害罪	抢劫罪	交通肇事罪	总样本
案件数	23300	6300	12835	1763	3457	47655
主审法官数	275	231	262	156	216	340
案件数/法官数(件)	84.7	27.3	49.0	11.3	16.0	140.2

[20] 参见前引[1],白建军文,第153页;前引[1],王越文,第77页以下;吴雨豪:《认罪认罚“从宽”裁量模式实证研究——基于部分城市醉酒型危险驾驶罪的定量研究》,《中外法学》2020年第5期。

[21] 例如,在诈骗、抢劫案件中常涉及有多个被告人的情况,由于每个被告人均有独立的量刑情节和量刑结果,故被单独提取作为一个独立的观测值。

2. 变量设置

本文的因变量是被告人获得的量刑结果，即其被判处的有期徒刑刑期；本文最关键的自变量是案件的主审法官。根据刑事诉讼法的规定，案件审理分为独任审判和合议庭审判两种。因此，对案件主审法官的提取分为两种情况：在独任审判中，唯一的审判员就是该案的主审法官；在合议庭审判中，对整个审判活动起主持和协调作用的审判长是该案的主审法官。基于这一原则，笔者通过计算机软件对每一份裁判文书的落款部分进行分析读取，提取出每个案件的主审法官。^{〔22〕} 根据前文表1所示的分布情况可以发现，340名主审法官中的大多数都参与了多个罪名案件的审理，这为我们检验量刑偏差在不同案件类型之间是否存在稳定性提供了可能。

根据2015年《最高人民法院关于完善人民法院司法责任制的若干意见》第5条，法院审理案件应当“实行随机分案为主、指定分案为辅的案件分配制度”。但不可否认的是，不同法官审理的刑事案件仍然存在区别。为了实现前文提及的在控制了量刑情节的基础上建构量刑集体经验的目标，针对每类案件，笔者在所有裁判文书中从基本犯罪事实和其他量刑情节两个方面，提取了所有可能影响量刑结果的法定情节和酌定情节作为统计分析的控制变量。这些变量的构成如表2所示。

表2 量刑情节控制变量体系

	基本犯罪事实	其他量刑情节
盗窃	是否入户盗窃、是否携带凶器盗窃、是否扒窃、数额（较大、巨大、特别巨大）	是否对特殊人群盗窃（残疾人、孤寡老人、丧失劳动能力人）、是否盗窃特定款物（救灾、抢险、防汛、优抚、扶贫、移民、救济款物）、是否造成严重后果（自杀、精神失常）
诈骗	数额（较大、巨大、特别巨大）	是否对不特定多数人实施诈骗、是否对特定款物实施诈骗（救灾、抢险、防汛、优抚、扶贫、移民、救济、医疗款物）、是否对特殊人群实施诈骗（残疾人、孤寡老人、丧失劳动能力人）、是否造成严重后果（自杀、精神失常）
故意伤害	轻伤结果、重伤结果、轻伤数量、重伤数量、是否手段残忍	是否涉黑恶势力，是否持枪支、管制刀具或其他凶器，是否民间矛盾激化而引发，是否激情、义愤所引发
抢劫	抢劫次数，伤亡结果（轻微伤、轻伤、重伤、死亡），数额（较大、巨大、特别巨大），是否入户抢劫，是否在公共交通工具上抢劫，是否抢劫银行或者其他金融机构，是否冒充军警人员抢劫，是否持枪抢劫，是否抢劫军用物资或者抢险、救灾、救济物资	是否涉黑恶势力、是否持凶器抢劫、是否手段残忍、是否转化型抢劫
交通肇事	重伤人数、死亡人数、是否逃逸、是否逃逸致人死亡	是否醉驾、毒驾、无证驾驶、超载、驾驶无牌号车辆，是否驾驶安全装置不全或者安全机件失灵的机动车辆，是否驾驶与准驾车型不符的车辆
共用变量	被害人过错、是否认罪认罚、未遂、从犯、自首、坦白、自愿认罪、立功、谅解、赔偿、和解协议、累犯、前科、犯罪对象特殊、程序因素（速裁程序、简易程序、普通程序）	

〔22〕 根据裁判文书制作的规范要求，每一份裁判文书的末尾都会注明独任法官或者合议庭每位审判人员的姓名。该部分是本文提取主审法官数据的来源。

（三）统计模型与方法

本文需要测量两个参数：第一，在不同案件的量刑结果差异中，有多大比例的差异来源于案件所分配的法官的不同。第二，针对每一位主审法官，在控制了相关量刑情节之后，以法官集体经验为参照系，在每一个罪中法官量刑偏离的具体数值是多少。为了完成上述两个任务，本文要运用特定的统计模型。

针对第一个问题，本文将运用多层线性模型（multilevel models），对法官个体不同所导致的量刑差异的比例进行估计。多层线性模型从上世纪80年代被统计学家发明以来，在教育、社会学和心理学等领域被创新性地应用。在刑事司法领域，也已经被犯罪学学者使用，用来测量法官个体差异对量刑结果的影响。^[23] 多层线性模型核心理念是，将多层结构数据在因变量上的总变异明确区分成组内和组间两个层次，然后分别在不同的层次上引入自变量来解释组内变异和组间变异。^[24] 以样本中的盗窃案件为例，我们关注的因变量是案件的量刑结果，其变异可以分为两个层次：第一，数据库中存在23300个盗窃案件，这是最底层的数据单位（层1），即组内分析单位。第二，这些案件被275位法官分别审理，每位法官审理的全部盗窃案件构成一个专门的数据组，案件的主审法官就是组间分析单位（层2）。在明晰了数据结构之后，就开始运用多层线性模型，通过以下步骤对参数进行估计：（1）通过使用最大似然估计的方法，获得模型的方差和协方差，^[25] 从而分解出被估计参数的变异构成；（2）将结果变量的方差分为组间方差和组内方差；（3）根据两部分的方差成分估计值，可以计算得出组内相关系数，以衡量层2单位之间的差异在层1结果变异的总方差中的影响。^[26]

因此，通过使用多层线性模型，能够将数据库中某一类刑事案件的量刑差异分解为两个部分：（1）案件本身所导致的量刑结果差别（组内方差 δ_e^2 ）；（2）案件分组所产生的量刑差别，也就是由不同法官审理所导致的量刑差异（组间方差 δ_j^2 ）。组内相关系数 ρ 的公式为：

$$\rho = \frac{\delta_j^2}{\delta_j^2 + \delta_e^2}$$

该系数刻画了，在不同被告人的量刑差异中，有多大比例是由审理法官的不同所导致的。该系数的值越小，意味着层2单位之间相对差异不大，即不同法官的量刑结果不具有显著差异。反之，该系数的值越大，意味着由不同法官审理而导致的量刑差异很大，法官的自由裁量权存在被滥用的风险。

在宏观上估算了法官自由裁量权对量刑差异的影响比例之后，还需要在微观上对每位法官的量刑偏离进行识别。由此涉及前述第二个参数的估计，即如何计算每位法官在每一种案件类型中的量刑偏离。为了实现量刑集体经验的测量，我们通过以下步骤完成参数估计：（1）以量刑结果为因变量，前文表2中的变量体系为自变量，建立量刑预测模型。（2）通过计算机循环程序，寻找在控制相关量刑情节之后量刑结果恰好等于集体经验的基准法官（在模型中

[23] See B. Kim, C. Spohn & E. C. Hedberg, *Federal Sentencing as a Complex Collaborative Process: Judges, Prosecutors, Judge - prosecutor Dyads, and Disparity in Sentencing*, 53 (4) *Criminology* 597 - 623 (2015).

[24] 参见前引[10]，谢宇书，第296页。

[25] 简单地说，方差和协方差估计的都是结果的波动性。前者衡量单变量的结果的波动性，比如某一特定法官在同类案件中的量刑差异大小。后者衡量多维变量中结果波动的一致性，比如某一位法官与其他法官的量刑一致性程度。

[26] 本文不是一篇介绍统计学方法的论文，对多层线性模型的计算步骤只作描述性的一般介绍。关于模型的具体计算方法，参见前引[10]，谢宇书，第294页以下。

称为“法官0”)。(3) 将案件的其他主审法官(在模型中称为“法官1”到“法官N”)作为自变量纳入模型,测算出在控制了量刑情节之后,每位法官相对于基准法官的量刑偏离值。由于基准法官的量刑结果恰好等于集体经验,所以,其余法官与这位法官的量刑均值偏离就是与集体经验的偏离。^[27]关于量刑偏离的统计模型,可以表示为如下形式:

$$y(\text{有期徒刑刑期(月)}) = \alpha_0 + \alpha_1 \text{量刑情节1} + \alpha_2 \text{量刑情节2} + \cdots + \alpha_n \text{量刑情节n} + \beta_1 \text{法官1} + \beta_2 \text{法官2} + \cdots + \beta_N \text{法官N}$$

这一多元线性回归模型的分析单位是每一个量刑决策。其中,因变量是该案中被告人被判处的有期徒刑刑期(月); α_0 是模型的截距项,其是指在所有的自变量均为0时,因变量的均值; α_1 到 α_n 是关于所有可能影响量刑的情节回归系数,它们与量刑情节变量一起构成了有关量刑情节的控制变量矩阵。法官1到法官N是一系列有关“案件由谁审理”的虚拟变量(只有0和1两个取值)。其中, β_1 到 β_N 刻画的是,相对于集体经验均值,在控制了相关量刑情节后,从法官1到法官N相对于集体经验均值的量刑偏离值。例如,如果 β_1 的值为3,意味着在控制了相关情节后,法官1的量刑均值比集体经验的量刑均值重3个月。

最后,由于各个罪名的量刑幅度不同,法官审理不同类型案件时的量刑偏离绝对值的变动幅度,也会发生相应变化。例如,盗窃罪基本犯的量刑幅度是3年以下有期徒刑,而抢劫罪基本犯的量刑幅度是3年至10年有期徒刑,由此,法官在抢劫案件中的量刑偏离绝对值一般会大于盗窃案件中的量刑偏离绝对值。为了使量刑偏离值在不同罪名之间能够横向比较,需要对上述量刑偏离值作标准化处理,即计算出它们的z值,其公式为:

$$z(\text{标准化的量刑偏离}) = \frac{\text{某法官在该类案件中的量刑偏离值} - \text{审理该类案件所有法官的量刑偏离均值}}{\text{审理该类案件所有法官量刑偏离的标准差}}$$

这一公式将每位法官的量刑偏离值与审理同一类案件其余法官的量刑偏离值进行比较,获得其偏离均值的标准差数量。例如,如果得出某一法官在某一类案件中量刑偏离的z值为2,就意味着,从平均意义上说,该法官的量刑偏离值高于所有法官量刑偏离值2个标准差。

四、量刑差异的宏观解构

运用多层线性模型,本文分别计算出每类刑事案件中量刑差异由案件本身所导致的部分和由审理法官不同所导致的部分,并以此对量刑差异进行解构。其结果如下文表3所示。以盗窃罪为例,案件本身所导致的量刑结果,组内方差估计值是214.2;审理法官不同所导致的量刑结果,组间方差估计值是18.08。基于上述结果,计算出一个组内相关系数 $\rho = 18.08 / (18.08 + 214.2) = 0.078$ 。这就是说,在北京地区盗窃案件的量刑差异中,案件层次量刑差异中大约有7.8%是由法官层次的差异造成的。关于其他案件类型,我们获得以下结论:在诈骗罪的量刑中,审理法官的不同可以解释6.1%的量刑差异;在故意伤害罪中,可以解释11.2%;在抢劫罪中,可以解释7%;在交通肇事罪中,可以解释5.3%。

从统计结果中,可以获得以下两个有关量刑差异宏观解构的实证结论。第一,“案件由谁

[27] 在实际的参数估计中,步骤2和步骤3交替循环进行,因为在一开始,我们并不知道哪位法官的量刑均值恰好等于集体均值,所以需要计算机依次对数据库中法官的量刑均值进行检验,直到找到那位量刑均值恰好等于集体均值的基准法官。

审理”能够解释相当一部分的案件量刑差异，法官自由裁量权对量刑结果的影响程度值得重视。统计结果发现，在北京地区审理的五类刑事案件中，案件审理法官不同所导致的量刑差异均在5%以上，在个别类型案件中，这一数值甚至达到10%以上。值得说明的是，这部分量刑差异均是案件被分配给不同法官所致，与案件本身的情况并不直接相关。举例而言，在不同法官的量刑差异可以达到10%的情况下，一个宣告刑在3年有期徒刑上下浮动的被告人，其可能由于案件被分配给不同的法官审理，而面临4个月左右的刑期浮动。从被告人的视角看，这4个月的刑期浮动与其犯罪事实和人身危险性无关，而仅仅与案件由谁审理有关，因此，该部分量刑结果具有相当的不可预知性。另一方面，通过将这里的数值与其他特定量刑情节进行比较，法官个体量刑差异过大所导致的后果将进一步显现。例如，量刑指导意见规定，对于当庭自愿认罪的被告人，“可以减少基准刑的10%以下”；对于有前科的被告人，“可以增加基准刑的10%以下”。如果法官自由裁量权所导致的量刑差异达到一定幅度，那么，体现在最终量刑结果上，该类量刑情节的实际效果就可能被消弭。例如，在不同法官的量刑差异可以达到10%的情况下，即使其他因素相同，却仅仅因为案件的主审法官不同，一个当庭认罪的被告人仍有可能被判处和另一个未当庭认罪的被告人一样甚至更重的刑期，一个有前科的被告人仍有可能被判处和另一个没有前科的被告人一样甚至更轻的刑期。这明显背离量刑指导意见所追求的量刑公正目标和罪刑相适应原则。

表3 多层线性模型结果与组内相关系数估计

	组间方差		组内方差		组内相关系数
	估计值	标准误	估计值	标准误	估计值
盗窃罪	18.08	4.25	214.20	14.64	7.78%
诈骗罪	125.30	11.19	1944.70	44.10	6.05%
故意伤害罪	23.14	4.81	182.34	13.50	11.26%
抢劫罪	105.10	10.25	1393.80	37.33	7.01%
交通肇事罪	7.65	2.77	137.22	11.71	5.28%

第二，法官个体导致的量刑差异幅度，在不同类型案件之间呈现出差异性。我们发现，在五类案件中，法官个体之间量刑差异最小的是交通肇事罪，在5%左右。一种可能的解释是，交通肇事罪作为典型的过失犯罪，不同案件的被告人在主观恶性和人身危险性上不存在显著差别。同时，大量交通肇事案件的犯罪事实比较简单，法官很可能仅依据犯罪造成的实际后果对量刑结果作相应调整。因此，在交通肇事案件的刑罚裁量中，法官自由裁量权的施展空间相对狭窄，从而导致法官之间量刑差异普遍较小。在盗窃、诈骗和抢劫案件的刑罚裁量中，法官之间的量刑差异虽然较交通肇事案件要大，但维持在较低水平——在6%到8%之间。其原因同样可能是，法官的自由裁量权容易受到特定量刑情节的规范约束。例如，在盗窃和诈骗这类纯粹财产犯罪案件的刑罚裁量中，犯罪数额对刑罚裁量起到决定性作用。在抢劫这样的人身财产犯罪复合型案件的刑罚裁量中，法官也多以涉案金额和被告人所受人身侵犯严重程度进行刑罚裁量，自由裁量权的空间同样受到严格约束。

我们发现，法官量刑个体差异最大的案件类型，是故意伤害这样的纯粹侵犯人身的暴力犯罪案件。一种可能的解释是，在故意伤害案件中，除了伤害结果这一相对客观的情节外，还存在大量缺乏客观标准的法定和酌定情节，这些情节的存在给法官量刑自由裁量权的充分发挥提

供了空间。例如，相当数量的故意伤害案件中存在错综复杂的案情缘由和情境因素，其中涉及被害人过错、家庭琐事、邻里纠纷、被告人冲动犯罪等多种因素的排列组合，这类因素均可能对量刑结果产生一定程度的影响。同时，由于生活经历和价值观不同，不同法官很可能对这些潜在的量刑情节存在不同的认知和评价，并以此行使自由裁量权，从而导致法官群体在故意伤害案件中呈现出最大的量刑差异性。

五、量刑偏差的微观识别

宏观部分的实证研究表明，在北京地区五类案件的刑事审判中，“案件由谁审理”能够在总体上解释相当一部分量刑差异。如前所述，法官自由裁量权导致的量刑差异过大，势必会对司法公信力产生冲击。同时，审理法官不同所导致的量刑结果不可预知性，也违背了基本的量刑公正理念。

因此，需要根据前文提出的“通过集体经验规制个体决策”的方案，对部分法官的量刑自由裁量权进行规制。具体而言，在对大多数与集体经验偏离不大的法官量刑决策采取肯定和尊重态度的同时，我们应当将目光聚焦于那些量刑显著偏离集体经验的少数法官。统计世界的经验法则表明，这些显著偏离集体经验的量刑结果理应属于小概率事件，因此，我们倾向于认为，在这些法官的量刑决策中很可能存在规范理解和价值判断上的偏差。通过纠正这些偏差，最终可以实现不断减少量刑畸重或者畸轻的情况，将法官群体的量刑自由裁量权控制在合理限度内的目标。

（一）单一案件类型的量刑偏差识别

通过前文“统计模型与方法”部分介绍的步骤，本文利用软件计算出每位法官在特定案件类型中相对于集体经验均值的量刑偏离值。值得说明的是，这里的量刑偏离值是在控制了所有可能的量刑情节之后，不同法官在个案审理中呈现出的量刑特征，因而被归为纯粹由自由裁量权所导致的量刑偏离。对量刑偏离值作标准化处理后，下文图2呈现了不同案件类型中法官量刑偏离的统计分布。

从图2可以发现，在所有案件类型中，法官的量刑偏离值均符合类正态分布。这意味着：第一，大多数法官的量刑结果均在集体经验均值附近。我们发现，不同案件类型中法官量刑偏离值分布的顶点的横坐标都在0附近——在控制了所有的量刑情节后，被告人最可能获得的量刑裁判结果就是接近集体经验均值的量刑值。同时，大多数法官的量刑偏离值都在距离集体经验均值1个标准差之内。这证明了我们的如下猜测：大多数法官的量刑结果与集体经验差别不大，因而不存在纠偏的必要。第二，有少部分法官的量刑结果与集体经验有较大偏离。在每一个案件类型的分布中，我们均发现，有一部分法官的量刑结果显著不同于法官群体的集体经验。其量刑结果或者显著重于集体经验均值（分布的右端），或者显著轻于集体经验均值（分布的左端）。但是，相较于在集体经验均值附近量刑的法官，这部分法官出现的频率显著较低。

因此，图2中的定性描述完全可以进一步证明如下论点：在量刑自由裁量权的规制中，我们要把目光聚焦于那些量刑畸重或者畸轻的少数法官，而对量刑结果在集体经验均值附近浮动的大多数法官的自由裁量权采取肯定和尊重的态度。但问题在于，如何确定这里的“少数法官”的范围。换言之，法官的量刑结果与集体经验偏离到何种程度时，才能被认定为量刑畸重或者畸轻？这一问题实际上涉及如何设定量刑自由裁量权的合理边界，而这又与量刑自由裁

量权规制的政策选择息息相关。当我们倾向于压缩法官量刑自由裁量权的空间时,就会将偏离的阈值调整到一个较低的水平,即法官的量刑结果与集体经验的偏离达到一个较小的数值时,就被认定是量刑畸重或者畸轻的征兆,从而有纠偏的必要。反之,如果我们倾向于赋予法官群体更多的量刑自由裁量权,就会将偏离的阈值调整到较高的水平,即只有当法官的量刑结果较大幅度地偏离集体经验均值时,才有纠偏的必要。

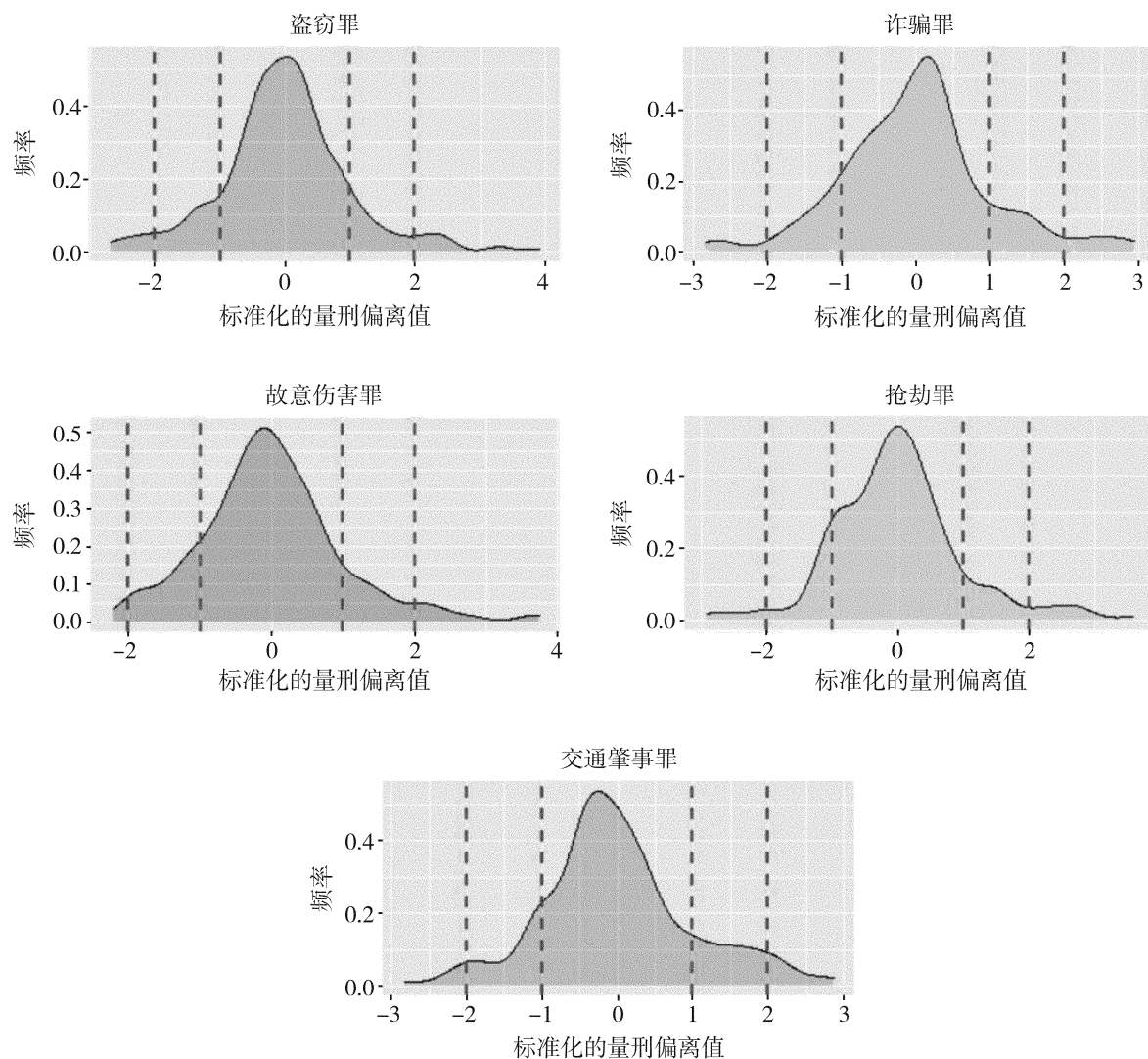


图2 不同案件类型的量刑偏离分布

根据上述两种政策选择,本文分别设定了两种不同的量刑偏差识别标准。由于法官的量刑偏离符合类正态分布,根据统计学中正态分布的经验法则,其中大多数的数值(超过68%)都应当分布在距离分布中心1个标准差之内的范围内;同时,更应当有绝大多数的数值(超过95%)分布在距离分布中心2个标准差之内的范围内。反之,如果数值超出上述范围,则被认定是相对的小概率事件。^[28] 将这一统计经验法则迁移到法官量刑偏差的识别中,本文认

[28] 参见[美]威廉·M. 门登霍尔:《统计学》,关静等译,机械工业出版社2018年版,第157页。

为：（1）当法官的量刑结果偏离集体经验达到1个标准差时，可以认为其量刑与集体经验存在较明显的差异。在倾向于压缩法官自由裁量权的政策选择下，该类量刑决策应当引起相应的重视。（2）当法官的量刑结果与集体经验的差距达到2个标准差时，法官的量刑已经显著背离集体经验，很可能超出了一般意义上合理行使自由裁量权的范畴。因此，即使在赋予法官更多自由裁量权的政策背景下，该类量刑决策也存在纠偏的必要。

表4 量刑偏差识别数据库的总体情况

	盗窃罪	诈骗罪	故意伤害罪	抢劫罪	交通肇事罪
主审法官数	275	231	262	156	216
量刑重于集体经验1个标准差以上的人数	34	32	39	18	28
量刑重于集体经验2个标准差以上的人数	11	10	14	8	10
量刑轻于集体经验1个标准差以上的人数	31	29	34	17	22
量刑轻于集体经验2个标准差以上的人数	9	4	9	4	5
进入偏差识别库的法官比例（以1个标准差为阈值）	23.64%	26.41%	27.86%	22.44%	23.15%
进入偏差识别库的法官比例（以2个标准差为阈值）	7.27%	6.06%	8.78%	7.69%	6.94%

本文根据上述两套阈值体系，构建了量刑偏差识别的数据库。事实上，在笔者构建的量刑偏差识别数据库中，显示有每位法官的姓名，以及其量刑结果与集体经验均值偏离的具体数值，从而能够实现个别化的量刑偏差识别及纠偏。出于隐私保护的考虑，本文不呈现这一数据库的具体内容。表4呈现的是北京地区法官量刑偏差识别数据库的总体情况，从中可以得出以下几点结论。

第一，大多数法官的量刑结果均在集体经验均值附近浮动。如果以量刑结果偏离集体经验达到1个标准差为阈值，在每个案件类型中，大约有22%—28%的法官进入量刑偏差识别库。这说明，四分之三左右法官的量刑结果都在集体经验均值的1个标准差范围内。如果以2个标准差为阈值，我们发现，每类案件均只有极少数法官进入量刑偏差识别库，所占比例在6%—9%之间。因此，我们的量刑偏差识别库能够实现将量刑自由裁量权的规制聚焦于少数法官的目的。

第二，在每一种案件类型中，无论采取何种阈值标准，量刑畸重的法官人数都显著多于量刑畸轻的人数。这说明，在量刑自由裁量权的规制中，关注的重点应当是那些践行重刑主义的法官群体。

第三，不同案件类型中进入量刑偏差识别库的法官比例存在差异。我们发现，无论采取哪一种阈值标准，在故意伤害案件的刑罚裁量中，进入量刑偏差识别库的法官比例都是最高的：以1个标准差为阈值，有近28%法官的量刑结果与集体经验存在明显偏离；以2个标准差为阈值，也有近9%的法官进入了量刑偏差识别库。同时，在故意伤害案件的刑罚裁量中，无论是量刑畸重、还是量刑畸轻的法官人数，都高于其他案件类型。这恰好印证了前文宏观实证研究部分的结论：由于复杂量刑机制的存在，在故意伤害案件中，法官的量刑自由裁量权有更大的发挥空间，量刑决策更可能呈现出较大的个体差异，因而更有规制和纠偏的必要。

（二）跨案件类型的量刑偏差识别

如前所述，如果法官仅在一种案件类型的刑罚裁量中存在偏差，那么，这很可能是因为该法官对某一犯罪的社会危害性、特定量刑情节的适用有不同于其他法官的理解。针对这种情

况,只需要对该法官某个罪名上的刑罚裁量行为予以纠偏。但是,一种更需要引起重视的情况是,某些法官在数个罪名的刑罚裁量中均出现量刑偏差。在这种情况下,该类法官很可能在量刑步骤和方法、刑罚裁量的价值判断乃至刑事政策的宏观理解上存在偏差。如果这种偏差不断持续和强化,势必会对司法公信力产生更加不利的影响。

因此,在这一部分,我们将对法官的量刑偏差进行跨案件类型的比较。在我们的数据库中,共有340位主审法官,其中大多数法官均审理超过一种类型的刑事案件。表5呈现了以1个标准差为偏差阈值的情况下,跨案件类型的量刑偏差识别情况。从该表中可以发现,有130位法官在一个罪名的刑罚裁量中存在偏差。由于数据库中有五种不同类型的犯罪案件,如果仅在一个罪名的刑罚裁量中存在偏差,并不能由此推断出法官在刑罚裁量的方法和理念上存在根本性偏误。与此对应的是,有73位法官在两个以上的罪名中均出现量刑偏差,有15位法官在三个以上的罪名中均出现量刑偏差,更有5位法官在四个以上的罪名中均出现量刑偏差。相对于仅在单一案件类型中出现量刑偏差的法官,后面几类法官的刑罚裁量偏差更具有跨案件类型的普遍性,因此更有必要规制其量刑自由裁量权。

表5 跨案件类型的量刑偏差识别情况

	总数	比例
仅在一个罪名中出现量刑偏差	130	38.24%
在两个以上罪名中出现量刑偏差	73	21.47%
在三个以上罪名中出现量刑偏差	15	4.41%
在四个以上罪名中出现量刑偏差	5	1.47%
主审法官数	340	

在多个案件类型中均出现量刑偏差,可能存在两种情形:一种是法官在各类案件中出现不同向量刑偏差,即其在某类案件中量刑畸重,在其他类案件中量刑畸轻。另一种是法官在各类案件中出现同向量刑偏差,即其在各类案件中均量刑畸重或者量刑畸轻,造成如此局面的原因更可能是法官在量刑的价值选择或者刑事政策的理解上存在显著偏差。如表6所示,在所有法官中,有27位法官在超过两个罪名的刑罚裁量中均出现量刑畸重的情况,有7位法官在超过三个罪名的刑罚裁量中均出现量刑畸重的情况,更有1位法官在超过四个罪名的刑罚裁量中均出现量刑畸重的情况。这些法官之所以在多个罪名中均出现重于集体经验的刑罚裁量,或许是因为他们存在一定的重刑主义倾向,即在宽严相济刑事政策的背景下,着重于将刑罚裁量向严厉的方向倾斜,从而出现量刑畸重的普遍情况。

表6 跨案件类型量刑畸重或者畸轻的识别

	总数	比例
在两个以上罪名中出现量刑畸重	27	7.94%
在三个以上罪名中出现量刑畸重	7	2.06%
在四个以上罪名中出现量刑畸重	1	0.29%
在两个以上罪名中出现量刑畸轻	22	6.47%
在三个以上罪名中出现量刑畸轻	3	0.88%
在四个以上罪名中出现量刑畸轻	0	0.00%
主审法官数	340	

另一方面，我们也发现有 22 位法官在超过两个罪名的刑罚裁量中均出现量刑畸轻的情况，有 3 位法官在超过三个罪名的刑罚裁量中均出现量刑畸轻的情况。虽然该类法官在绝对人数上均少于量刑畸重的法官人数，但这同样说明，该类法官在刑事政策的理解上，致力于将刑罚裁量向轻缓化的方向倾斜。

六、大数据时代下量刑自由裁量权的规制

本文从近 5 万份北京地区基层法院审理的刑事案件的裁判文书切入，从宏观、微观两个层面提出了对特定地区法官群体量刑自由裁量权的规制方案。特别值得说明的是，基于特定研究目的的考虑，本文将数据样本聚焦于北京地区的刑事案件。但是，在全国范围内裁判文书公开的背景下，本文所用的方法可以无障碍地应用于北京以外的其他地区。更为重要的是，在本文实证研究的背后，实则蕴含着一个具有普适性的理论问题：在量刑规范化被列为重要司法改革项目的今天，如何通过运用大数据和人工智能，完成对量刑自由裁量权的合理制衡。

（一）量刑规范化与自由裁量权的衡平

在刑事案件的裁判中，量刑结果是否公正是案件裁判是否正义的最直接体现。正是在此背景下，量刑规范化成为近年我国司法改革的一个重要议题。其背后的价值理念是，通过顶层设计和具体规则设定，限制法官的量刑自由裁量权，缩减法官滥用权力的可能性，并最终实现量刑确定、一致和公平的价值目标。^[29]

一方面，量刑规范化改革及其价值追求具有充分合理性。在世界范围内，限制法官自由裁量权以实现量刑规范化，具有悠久的历史 and 广泛的根基。^[30] 在“相同的案件应当获得相同的裁判结果”理念的指引下，减少法官个人因素的影响，提高量刑结果的可预测性，符合人们最朴素的正义观念。

另一方面，对法官自由裁量权的控制必须要有一定的限度。试图以量刑规范化完全消除法官的量刑自主决策权，既不合理，也没有可行性。这是因为，量刑并不是单一、静态的机械计算行为，相反，其应当是多维的、动态的综合决策行为。在实践中，法官面对的是一个具体的案件和被告人，其在作出最终的量刑决定之前，需要在一系列可能影响量刑的因素中（例如责任、预防、刑事政策、社会关系的修复等）斟酌权衡，不可避免地会融入自身的价值判断。^[31] 由于这种复杂机制的存在，我们应当给予法官一定的行使量刑自由裁量权的空间，允许其在个案中作出符合其良心的公正的量刑裁判。即使这种公正裁判的判断标准在不同量刑主体之间存在差异，但只要不超出合理限度，就应当被认定为正当。反之，如果一味试图压缩法官的量刑自由裁量权，并否定个体价值判断在量刑裁判中的合理空间，就会导致量刑裁判走向形式化、机械化。

因此，在肯定量刑规范化改革积极方面的同时，也要尊重法官在刑罚裁量过程中的自由裁量权。事实上，这也是学者所言的“量刑规范化与法官量刑自由裁量权的衡平”。^[32] 然而，

[29] 参见熊秋红：《中国量刑改革：理论、规范与经验》，《法学家》2011年第5期，第46页。

[30] 参见李洁、于雪婷、徐安怀：《量刑规范化的规范方式选择》，《当代法学》2011年第3期，第69页。

[31] 参见前引〔3〕，彭文华文，第110页。

[32] 参见臧冬斌：《量刑规范化与法官量刑自由裁量权的衡平》，《河北法学》2007年第12期，第113页。

从提出一种理念的角度,对两种相互冲突的取向进行衡平,固然能够得到普适性的答案,但从实际操作的角度看,受制于研究方法的限制,既有的研究和规则制定并没有很好地解决上述衡平中可能存在的“二律背反”问题。

本文通过特定的定量研究方法,设计出了一套具有可操作性的规制法官量刑自由裁量权的方法。具体而言,与以往“一刀切”地修订量刑指导意见或者颁布量刑实施细则不同,本文提出的对量刑自由裁量权的制衡方法采取的是一种“个别化”的方案。我们将目光聚焦于特定罪名的刑罚裁量和特定法官的量刑偏差识别,只有当法官群体在某一案件类型中的量刑差异达到警戒值,或者法官个人的量刑偏离超过合理限度时,对量刑自由裁量权的规制才有必要;反之,则应充分尊重法官的量刑决策。如下文图3所示,上述量刑自由裁量权规制方案分为以下几个步骤。

首先,监督者需要在宏观上对量刑差异进行分析解构,从而发现量刑自由裁量权规制的潜在问题。这是量刑自由裁量权规制的基础步骤。当监督者发现,在某一类案件中,“案件由谁审理”较大幅度地影响刑罚裁量结果时,其需要警惕,在这类案件的审理中,可能有必要控制法官的自由裁量权。例如,本文通过北京地区五个常见罪名案件量刑样本的数据分析发现,“案件由谁审理”能够解释超过5%的量刑差异;其中,在故意伤害案件的量刑中,这一差异达到了11%。如果监督者认为上述结果已经超出合理边界,就需要对法官的量刑自由裁量权进行相应控制。

其次,监督者需要在微观上对特定法官的量刑偏差进行识别。控制量刑自由裁量权的关键,是以集体经验为参照系,缩小量刑的离散程度或者量刑偏差,减少极端个案。^{〔33〕}因此,通过运用特定统计方法计算出每位法官的量刑偏离值之后,监督者需要找出那些在刑罚裁量中容易出现极端个案的法官群体。这意味着,一方面,监督者要对量刑结果未显著偏离集体经验的法官群体的量刑决策采取肯定和尊重的态度;但另一方面,基于量刑规制的政策选择和统计经验法则,监督者需要找出那些量刑畸重或者畸轻的法官。这类法官的量刑决策,是量刑规范化改革中需要着重关注的。例如,我们的实证研究发现,如果以2个标准差的量刑偏离作为自由裁量权规制的阈值,将会有6%—9%的法官进入量刑偏差识别库。这也意味着,只需要重点关注这些法官的量刑决策,而对剩余90%以上法官的量刑决策采取肯定和尊重的态度。

最后,监督者需要对特定法官的量刑偏差作归因解释。法官的量刑偏差可能由多种原因导致,包括法官对量刑情节的理解偏差、对量刑步骤和方法的理解偏误以及刑事政策价值选择上的偏离。只有完成对特定法官量刑偏差的归因解释,量刑偏差的纠正才能实现对症下药。本文提出了一种依定量方法进行量刑偏差归因解释的方法:对法官的量刑偏差进行跨案件类型的比较,以发现法官是否在数个罪名的刑罚裁量中均出现量刑畸重或者畸轻的情况。例如,有7位法官在三个以上罪名的刑罚裁量中均出现量刑畸重的情况,我们倾向于将这些法官的量刑偏差归因于他们在刑事政策上普遍采取重刑主义的立场。

应当承认的是,量刑偏差的归因是一个复杂的系统性工程,定量研究不足以完成所有解释任务。在此之外,还需要采用对法官进行访谈、调查的方法,通过定性研究探寻出现量刑偏差的潜在机制。

〔33〕 参见前引〔1〕,白建军文,第153页。

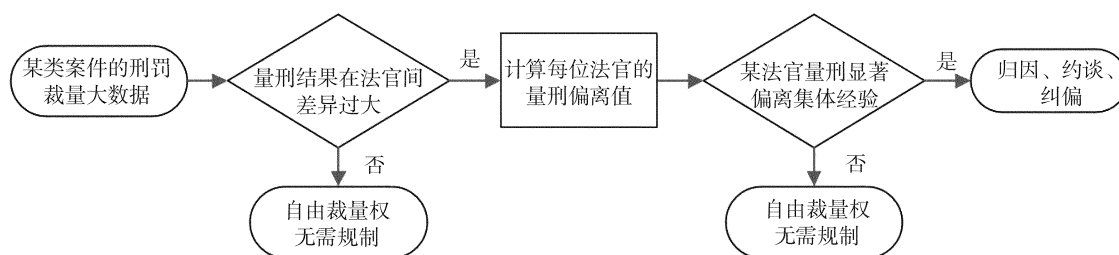


图3 量刑自由裁量权的规制流程

（二）大数据与算法在量刑规范化改革中的应用

实施前述量刑自由裁量权规制方案的前提，是存在反映量刑集体经验的大样本。近年来，裁判文书公开制度为我们从司法大数据中归纳、总结司法裁判集体经验提供了可能。自2013年中国裁判文书网上线以来，我国裁判文书公开上网实现了巨大进步。2016年《最高人民法院关于人民法院在互联网公布裁判文书的规定》更是明确提出，裁判文书“以公开为原则、不公开为例外”。^[34]随着海量文书的上网，一方面，公开的每一份裁判文书都有可能受到社会大众的检阅。案件当事人和法律专业人士能够从以往的个案中总结裁判规则，并对司法实践进行监督。另一方面，数量众多的裁判文书以一种较为完整的方式提供了各级法院的司法审判图景。对海量裁判文书进行数据化处理后，就能获得可供分析的司法裁判大数据，从而成为法律实证研究的富矿。

在本文的研究过程中，笔者通过一定的技术手段，对2014—2020年北京地区公开的所有基层法院审理的五类刑事案件裁判文书作了有目标的提取与分析。例如，基于量刑自由裁量权这一研究主题，在数据提取过程中，笔者着重关注“案件的审理法官”这一变量，并将其与案件的刑罚裁量结果进行关联，从而为利用大数据展开量刑自由裁量权规制和量刑偏差识别提供了可能。在获得数据后，笔者进一步运用特定算法对数据进行分析，并最终得出了可供决策者参考的结论。

由此可见，在技术层面，本文提出的量刑自由裁量权规制方案的实施，包括了“数据提取—算法应用—结论参考”的过程。并且，上述算法一旦开发完成，即可在不同地区、不同时间重复使用。事实上，在本文写作过程中，笔者基于该方案开发了一套量刑偏差识别程序，其完整地包括了上述从数据到算法再到结论的全部过程。使用者只要在程序中导入某一地区法院在某一时间范围内的所有刑事裁判文书，该程序就会基于本文介绍的算法，自动输出不同罪名案件中法官个人因素所导致的量刑差异、不同法官的量刑偏离值以及可能具有量刑纠偏必要的法官名录，从而能够为量刑自由裁量权的规制提供重要的决策参考。

应当说，与传统的监督者基于个案发现并纠正量刑偏差的方案不同，这一通过机器和算法进行量刑自由裁量权控制的方案，有其特定优势。第一，能够在短时间内完成对大量案件、大量法官刑罚裁量的分析和识别，快速完成定量计算并生成结果，从而保证了监督的高效性。^[35]第二，在量刑偏差的识别过程中，能够运用特定算法控制住所有可能影响刑罚裁量结果的其他

[34] 参见马超、于晓虹、何海波：《大数据分析：中国司法裁判文书上网公开报告》，《中国法律评论》2016年第4期，第196页。

[35] See Christopher Rigano, *Using Artificial Intelligence to Address Criminal Justice Needs*, National Institute of Justice Journal, Issue 280, 2019, p. 2.

量刑情节,从而计算出纯粹由法官个人因素导致的量刑偏离,进而保证偏差识别的相对准确性。第三,由于机器和算法不受人的主观情感控制,保证了对所有案件、所有法官均以同一套标准进行评价,从而在一定程度上保证了监督和决策的公正性。^[36]事实上,在司法实践中乃至社会管理领域内,引入大数据和人工智能以辅助特定主体进行决策,在一些国家已经被广泛运用。^[37]

但是,我们也要清醒地意识到,在任何时候,机器都不可能完全代替人类作出主观决策和价值判断。在司法实践中引入算法作出决策,也面临特定风险。^[38]这种风险体现在,算法很可能存在失误和不透明的情况,很可能不断重复错误的决策;在特定情形下,算法的使用还可能导致结构化歧视。^[39]

在量刑自由裁量权规制过程中,在预设量刑决策者对量刑结果有完全决定能力的前提下,我们将量刑结果相对于集体经验的偏差作为评价个体决策的依据。然而,在现实世界中,这一前提并不一定总是得到满足。一方面,虽然案件的主审法官对量刑结果有最终决定权,但不可否认的是,在合议庭审判的情况下,合议庭其他审判人员的意见仍然可能对量刑结果产生影响。另一方面,虽然审判权的独立公正行使、主审法官对案件负责,是近年来司法改革的重要方向,但现实中,法官作为司法体制运行中的一员,其个体决策仍然可能受到系统内其他人员的影响。例如,在审判委员会、法官的上级领导对案件发表意见时,法官量刑决策的自主性就可能相应减弱。

应当承认的是,上述问题难以为实证研究和算法所完全揭示和克服。因此,对算法在量刑规范化改革中的作用,应保持相对谨慎的态度。例如,国际上算法伦理学者提出了“算法的程序控制”理念,用以规避算法在决策中产生的道德风险。其核心观点是,在算法代替人类进行决策的同时,必须在其被使用的各个阶段受到人类的监督和纠正,从而保证算法的公正:(1) 算法依据的数据必须透明并且受到检验;(2) 算法依据的计算方法必须公开和正确;(3) 被算法监督者必须要对算法的监督知情;(4) 算法不能成为决策的唯一依据。^[40]

将上述理论框架迁移到利用裁判文书数据规制量刑自由裁量权的领域,本文认为,一方面,在算法建构的技术层面,要对算法所依据的数据和方法保持科学严谨的态度。例如,各地法院应不断提高司法透明度,提高裁判文书的公开化水平,尽力避免选择性公开的情况。^[41]在具体模型的选择上,监督者必须及时公开其依据的算法,从而使统计学者、法律学者以及实务工作者能够对算法的科学性和准确性作出适时评估。^[42]算法开发者在收到反馈意见后,应

[36] See V. Chiao, *Fairness, Accountability and Transparency: Notes on Algorithmic Decision-making in Criminal Justice*, 15 (2) *International Journal of Law in Context* 138 (2019).

[37] See P. D. König & T. D. Krafft, *Evaluating the Evidence in Algorithmic Evidence-based Decision-making: the Case of US Pretrial Risk Assessment Tools*, 33 (3) *Current Issues in Criminal Justice* 359 – 360 (2021).

[38] See S. Levmore & F. Fagan, *Competing Algorithms for Law: Sentencing, Admissions, and Employment*, 88 (2) *University of Chicago Law Review* 406 – 408 (2021).

[39] See K. Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates*, 23 (4) *Theoretical Criminology* 453 – 470 (2019).

[40] See D. K. Citron & F. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 (1) *Washington Law Review* 1 – 34 (2014).

[41] See B. L. Liebman, M. E. Roberts, R. E. Stern & A. Z. Wang, *Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law*, 8 (2) *Journal of Law and Courts* 197 – 198 (2020).

[42] 例如,本文在研究方法部分详细介绍了宏观量刑差异和微观量刑偏差的计算和识别方法。

对算法进行相应修正，从而保证算法在科学性、高效性和实用性方面不断完善。

另一方面，在算法的具体运用上，要对其评价结果持谨慎态度。算法可以作为监督量刑自由裁量权、评价法官量刑行为的决策参考，但不应成为决策的唯一依据。在一定的时空范围内，基于集体经验的算法是统一的、静止的，但基于个体决策的量刑实践却永远是分化的、动态的。^{〔43〕}同时，由于量刑决策可能受到其他因素的影响，我们无法期待算法能够准确描绘和解释所有的现实问题，^{〔44〕}所以，在监督的过程中，不能一味对由算法识别出的量刑偏离作否定性评价。相反，算法的识别只是第一步，还需要通过访谈和实地调研等方式，并加入监督者的主观评价，在对量刑偏差进行全面归因解释后，再对法官的刑罚裁量作出公允评价。另外，法官作为刑罚裁量主体，必须对其量刑行为受到算法的监督这一事实知情。如果法官意识到算法的监督存在问题，其有权提出异议，并对算法的修正提出反馈意见。

Abstract: In the process of sentencing normalization reform, one issue that has drawn the attention of both academic and practical circles is how to acknowledge judicial discretion while keeping sentencing deviations within a legitimate limit. A feasible solution is to compare a judge's individual sentencing decision with the collective experience of judges and to adopt an affirmative and respectful attitude towards the discretion of judges whose decisions are close to the collective experience while identifying and correcting the decisions of judges that significantly deviate from the collective experience. This paper is an empirical study of nearly 50,000 criminal cases adjudicated in the People's Courts in Beijing, covering larceny, fraud, robbery, intentional injury, and traffic crimes. Firstly, the results of the multilevel model show that the variable "who hears and decides the case" can explain 5% – 11% of the sentencing disparity. Secondly, most judges' sentencing decisions fluctuate around the collective experience, and only a small portion of judges' sentencing decisions deviate significantly from the collective experience and, therefore, need to be corrected. Finally, after comparing sentencing deviations across different types of criminal cases, we found that a small number of judges have sentencing deviations in multiple crimes. These judges should become the targets of discretionary governance. This discretionary governance scheme demonstrates the application of big data and algorithms in sentencing normalization reforms and has the prospect of being popularized under the premise of introducing the concept of "due process for algorithms".

Key Words: sentencing normalization, judicial discretion, collective experience, deviation identification

〔43〕 参见前引〔38〕，Levmore等文，第411页。

〔44〕 参见李训虎：《刑事司法人工智能的包容性规制》，《中国社会科学》2021年第2期，第50页。