## Book Section:

# Which Are the Main Characteristics Determining Sentence Severity?

## An Empirical Exploration of Shoplifting Offences Using Spike-and-Slab Models

**Jose Pina-Sánchez · Mandeep K. Dhami · John Paul Gosling**

**Abstract** In order to identify the case characteristics considered by judges in their sentencing deliberations, researchers have relied heavily on regression modelling techniques. In most instances, the sample of sentences available is many times larger than the number of case characteristics recorded, providing enough degrees of freedom to estimate the effect of these case characteristics adequately. However, when the number of case characteristics recorded is too large, or the samples are too small, regression models can overfit the data. Here, we demonstrate how, in such settings, Bayesian methods such as spike-and-slab models can be used to select the most consequential case characteristics in a principled and reliable way. The potential of this approach is illustrated using a sample of 2,116 sentences imposed in the magistrates' courts in England and Wales on shoplifters. For this small sample, we reliably estimated twenty case characteristics predicting custody decisions. This highlights the high degree of discretion afforded to sentencers in England and Wales. We also found that offender-related factors (such as the offender's previous convictions, and caring responsibilities), appeared to be far more important than characteristics defining the offence (e.g., value of goods stolen or lasting effect of offence on victim). This questions the widely held belief that sentencing in England and Wales is based around the principle of proportionality.

Jose Pina-Sánchez
School of Law, University of Leeds, UK
Liberty Building, University Western Campus, Moorland Rd, Leeds LS3 1DB, UK
ORCID: 0000-0002-9416-6022
E-mail: j.pinasanchez@leeds.ac.uk

Mandeep K. Dhami
Department of Psychology, Middlesex University (London), UK
ORCID: 0000-0001-6157-3142

John Paul Gosling
School of Mathematics, Durham University, UK
ORCID: 0000-0002-4072-3022

## 1 Introduction

Judicial decision-making, and sentencing in particular, are complex cognitive processes. The number of case characteristics that could be deemed relevant and therefore influence the sentence outcome is often seen as unlimited, since, as sentencers like to say, *'no two cases are alike'*. Past attempts to empirically describe that process have been hampered by data limitations (Dhami and Belton, 2015), with most jurisdictions only making available a few of the main characteristics defining a criminal case (e.g. offence type, number of previous convictions, or whether the offender plead guilty). In the absence of detailed case-level data, key questions in the field of sentencing research remain only partially answered, and these answers likely biased (Dhami et al., 2004). For example, if we look at studies exploring gender disparities in sentencing; how do we know that apparent harsher treatment of male offenders is not due to female offenders' lower recidivist rates (Langan and Levin, 2002), likely reflected in pre-sentence reports recommending community sanctions? Similarly, how do we know that the effect attributed to previous convictions is not confounded by perceptions of offenders' dangerousness left uncontrolled in the analysis?

In 2014, the Sentencing Council for England and Wales changed the research landscape in that jurisdiction through the publication of data obtained from the Crown Court Sentencing Survey. This survey captured practically all of the factors listed in the sentencing guidelines (i.e. factors explicitly identified as relevant for sentencing specific offences), and led to an explosion of empirical sentencing research in the UK. As anticipated (Roberts, 2013), researchers used this data to estimate the effect of multiple contentious factors, which up to then had only been debated based on anecdotal or qualitative evidence. For example, new studies examined the aggravating effect of alcohol intoxication Lightowlers (2019), the effect of an offender's role in drug trafficking (Fleetwood et al., 2015), show of remorse (Maslen, 2015), caring responsibilities (Kane and Minson, 2022), guilty pleas (Pina-Sánchez and Gosling, 2020), or being charged with multiple offences (Dhami, 2022), to name a few.

However, the availability of more detailed sentencing data has also brought about new methodological challenges. One of them being the difficulty of controlling for multiple relevant case characteristics simultaneously, which is especially challenging when the sample size is not large. So far, sentencing researchers facing this problem have either: i) ignored it, throwing all recorded case characteristics in their statistical model; ii) decided themselves the most important case characteristics that ought to be selected (e.g., based on past research, theory, or current sentencing policy); or iii) used stepwise regression models to undertake the selection process. Ignoring the problem leads to overfitted models that in turn tend to be prone to multicollinearity (i.e. a model's incapacity to disentangle the specific effects of two or more predictors that are highly correlated). Rationales commonly used to select variables can lead to arbitrary decisions when the variables that end up being discarded are nevertheless relevant (i.e. considered by the sentencer). Whereas by unduly undertaking multiple significance tests, stepwise methods lead to underestimated measures of uncertainty (i.e. standard errors, confidence intervals and p-values are biased towards zero) (Harrell, 2001; Smith, 2018).

In order to identify the most consequential case characteristics when the sample size is relatively small, we suggest the adoption of spike-and-slab models (Forte et al., 2018; Ročková and George, 2014). In essence, these are a type of Bayesian selection method that uses a mixture prior with one component that concentrates mass at zero (the 'spike') and another that has a flat, diffuse distribution (the 'slab'). This allows spike-and-slab models to perform shrinkage and conduct variable selection within a single unified framework. In so doing, these models can overcome the trade-off arising from having to accept either overfitted models or following arbitrary decisions to limit the number of case characteristics to be selected.

Here, we apply these models to a new dataset published by the Sentencing Council for England and Wales, containing details of cases sentenced in the magistrates' court for theft from a shop (shoplifting). These represent an offence type and court level that, in spite of their volume and relevance,

remain importantly under-researched.[1] We show how spike-and-slab models can help us recognize the case characteristics that more clearly predict magistrates' sentencing decisions. Therefore, this study provides both a novel research tool to explore how different case characteristics are weighed in the sentencing process, as well as new empirical insights into the sentencing of shoplifting offences in the magistrates' court. Beyond its academic merit, the empirical contribution presented here also has an important policy application. This is because when developing sentencing guidelines the Sentencing Council for England and Wales must reflect current sentencing practice, for which robust empirical methodologies that yield valid and reliable findings are of the essence.

The next section provides more context into how current guidelines structure the sentencing of shoplifting offences in England and Wales. This is followed by sections describing the data we will model, our analytical strategy, findings, and their implications.

## 2 Jurisdictional Context: Sentencing Shoplifting Offences in England and Wales

Theft from a shop (or stall), colloquially known as shoplifting, is the highest volume offence in the 'Theft Offences Definitive Guideline' (Sentencing Council, 2015b). This guideline was published the 1st of February 2016, replacing an older version created by the Sentencing Guidelines Council.[2] One of the key differences of the new sets of guidelines introduced by the Sentencing Council is the adoption of a structure of nine sequential and non-overlapping steps in the decision-making process (Dhami, 2013), of which the first two are most critical.

At 'Step One' sentencers must determine the harm caused to the victim and the offender's culpability, both of which are classified into one of three categories. To do so the guidelines provide a comprehensive list of factors that sentencers ought to take into consideration. For example, the degree of planning or sophistication involved in the offence is used to determine the level of culpability, while the value of the goods stolen is used to define the level of harm (see Table 1 in Section 3 for the full list of Step One factors). For each combination of harm and culpability a specific starting point and range of 'appropriate' sentence outcomes are assigned.

In 'Step Two' sentencers must choose a preliminary sentence within the defined sentencing range. Here, sentencers may 'fine-tune' the starting point defined in Step One by considering further aggravating and mitigating factors. These factors do not form the principal factual elements of the offence, but instead provide the context of the offence and the offender. For example, sentencers may consider the presence of previous convictions, whether the offence was committed on bail, or whether the offender showed genuine remorse. Importantly, while the list of Step One factors provided in the guideline is comprehensive, those in Step Two are not, thus allowing sentencers to consider aggravating and mitigating factors outside the specific remit of the guidelines.

The preliminary sentence determined in Step Two can then be modified going through the additional sequence of steps. For instance, these include considering questions such as guilty pleas, time spent on remand, or whether the totality principle is applied (in multiple-offence cases).

---

[1] The magistrates' court process roughly 90% of criminal cases in England and Wales (Sturge, 2021).

[2] The design of the new theft guideline was informed empirically through the Sentencing Council's own research efforts. This involved content analysis of sentence transcripts, qualitative interviews and vignette designs with magistrates and judges (Sentencing Council, 2015c). From that study a number of differences between the two guidelines were identified, together with a potential increase in severity following the coming into force of the new guideline. The latter was corroborated in various impact evaluations (Sentencing Council, 2019; Pina-Sánchez et al., 2019), which showed a stronger process of sentence inflation than expected. A series of explanations have been put forward for this. Such as the new emphasis placed by the new guideline on the value component used to determine harm, or the higher saliency of previous convictions, which was consistently cited by magistrates as the most important factor consider when sentencing theft offenders (Sentencing Council, 2019). Other than those two factors, evaluations of the guidelines have pointed at the majority of case characteristics being applied as they used to be under the previous guidelines, and no major changes in consistency were detected (Isaac et al., 2021).

## 3 Sentencing Data: Shoplifters in the Magistrates' Court

We use data collected by the Sentencing Council for England and Wales from a sample of magistrates' courts in 2016. The original aim of the data collection exercise was to help the Council evaluate the impact of their guidelines. The data was then made available to the public in 2021.[3] Specifically, we use a sample of 2,116 cases of shoplifting sentenced under the 2016 sentencing guideline.

The data was obtained using self-completed questionnaires delivered to magistrates and district judges from 79 magistrates' courts.[4] These sentencers were instructed to fill out the questionnaires after sentencing an offender whose principal offence was theft from a shop or stall. In the questionnaire participants were asked to give detailed information on the factors they took into account when sentencing the offender, and on the final sentence they meted out.

Having the data directly provided by sentencers shortly after passing a sentence is a unique feature of the datasets produced by the Sentencing Council, particularly because the questionnaire covers most of the factors listed in the sentencing guidelines. This contrasts with typical sentencing studies relying on administrative data, which provide no more than a few factual case characteristics, such as offence type, guilty plea, bail, and previous convictions. The response rate for this specific survey is not reported, however, previous surveys of sentencers by the Sentencing Council have achieved a reasonably good response rate of over 60% (Sentencing Council, 2015a).

In spite of our ambition to explore *all* sentencing factors considered in the guideline, some of them are quite rare and were only seen in a few cases. Hence, we focus on case characteristics that are present in at least 1% of cases in our sample. Descriptive statistics of all the variables considered in our analysis are reported in Table 1. The most important of them is the final sentence, a multi-categorical variable that we dichotomise into 'suspended/immediate custody' or 'not' ('custody' hereafter). This binary variable serves as simple proxy for sentence severity, capturing whether the offender was sentenced to any of the two possible custodial sentences used in England and Wales or to an alternative, more lenient (non-custodial) disposal type, including discharge, fine, or community order. Cases that were sent by the magistrates' court to the Crown Court (a higher tier court) were excluded from our analysis since they were not sentenced in the lower tier court of interest.

The explanatory variables in our analysis can be classified into four groups according to whether they relate to culpability, harm, aggravation, mitigation, and offender characteristics. Most of these variables are binary, but amongst them there are also five ordinal variables: value of goods (1: 'Up to £10', 2: '£11-£50', 3: '£51-£100', 4: '£101-£200', 5: '£201-£500', 6: '£501-£1000', 7: '>£1001'), role of offender (1: 'lone', 2: 'limited', 3: 'significant', 4: 'leading'), use of force (1: 'none', 2: 'limited', 3: 'high'), level of planning (1: 'none/little', 2: 'some', 3: 'high'), and age (1: '18 to 21', 2: '22 to 29', 3: '30 to 39', 4: '40 to 49', 5: '50 to 59'). To facilitate the computation process for our models, these ordinal variables are centred around the mean and introduced as continuous variables.

As shown in Table 1, our analytical sample is composed of mainly male offenders, 72%, which is nonetheless a lower proportion than for most other offence types. The most common age category is 30 to 39, representing 38% of cases. And although there is no information on the ethnicity of the offender, we note that in 2013, 85% of shoplifting offenders sentenced in England and Wales were perceived to be of White origin by the police officer who handled their case (Sentencing Council, 2014).

---

[3] The data is available here, https://www.sentencingcouncil.org.uk/research-and-resources/data-collections/magistrates-courts-data-collections/theft-from-a-shop-or-stall/.

[4] The questionnaire is available here, https://www.sentencingcouncil.org.uk/wp-content/uploads/Post-guideline-data-collection-form-Theft-from-a-shop-or-stall.pdf.

**Table 1** Descriptive statistics of the variables used in the analysis

| Variable name | Mean | Min. | Max. |
|---|---|---|---|
| Sentence: suspended/immediate custody | 0.29 | 0 | 1 |
| Culpability: level of planning | 1.38 | 1 | 3 |
| Culpability: use of force | 1.06 | 1 | 3 |
| Culpability: role of offender | 1.37 | 1 | 3 |
| Culpability: sophisticated offence | 0.02 | 0 | 1 |
| Culpability: subject to a banning order | 0.01 | 0 | 1 |
| Culpability: involvement of others | 0.02 | 0 | 1 |
| Culpability: mental disorder | 0.05 | 0 | 1 |
| Harm: value of goods stolen | 3.37 | 1 | 7 |
| Harm: emotional distress | 0.01 | 0 | 1 |
| Harm: injury to victim | 0.10 | 0 | 1 |
| Harm: effect on business | 0.38 | 0 | 1 |
| Harm: other factors | 0.06 | 0 | 1 |
| Aggravating: previous convictions | 0.84 | 0 | 1 |
| Aggravating: conceal evidence | 0.07 | 0 | 1 |
| Aggravating: failure to comply | 0.22 | 0 | 1 |
| Aggravating: offender on bail | 0.17 | 0 | 1 |
| Aggravating: offences into consideration | 0.12 | 0 | 1 |
| Aggravating: harm to the community | 0.20 | 0 | 1 |
| Aggravating: professional offending | 0.04 | 0 | 1 |
| Aggravating: stealing goods to order | 0.03 | 0 | 1 |
| Aggravating: other factors | 0.25 | 0 | 1 |
| Mitigating: age / lack of maturity | 0.03 | 0 | 1 |
| Mitigating: good character | 0.02 | 0 | 1 |
| Mitigating: no recent convictions | 0.08 | 0 | 1 |
| Mitigating: financial hardship | 0.10 | 0 | 1 |
| Mitigating: steps to address addiction | 0.13 | 0 | 1 |
| Mitigating: mental disorder | 0.09 | 0 | 1 |
| Mitigating: remorse | 0.16 | 0 | 1 |
| Mitigating: return of stolen property | 0.02 | 0 | 1 |
| Mitigating: serious medical condition | 0.03 | 0 | 1 |
| Mitigating: primary carer | 0.02 | 0 | 1 |
| Mitigating: other factors | 0.15 | 0 | 1 |
| Offender: age band | 3.10 | 1 | 6 |
| Offender: male | 0.72 | 0 | 1 |

Note 1: The mean of the binary variables represents their proportion.
Note 2: Continuous variables are shown here in their original scale, but they were centred in our models.

## 4 Analytical Strategy: Spike-and-slab models

Our analytical strategy is based on the estimation of a series of spike-and-slab models, a type of Bayesian averaging technique (Mitchell and Beauchamp, 1998). These type of models, allow researchers to circumvent the arbitrariness associated with having to select one set of predictors when various sets could be valid, and provide unbiased uncertainty estimates.

A wide range of approaches has been suggested to undertake model selection in a principled way. For example, best subset selection finds the optimal model for each number of predictors through exhaustive search, but becomes computationally infeasible for a large number of predictors. This problem is circumvented through other frequentist approaches like LASSO, which perform shrinkage and variable selection by penalising model complexity in the optimisation objective. However, none of those methods provide an adequate characterisation of the uncertainty of the model selection process.

Bayesian approaches take a more direct approach to accounting for model uncertainty, calculating the posterior probability of all possible models and averaging predictions across models weighted by their posterior probability. Specifically, for the case of spike-and-slab models, the selection process is facilitated through a Bernouilli mixture prior distribution composed of a Dirac delta function (the 'spike') and a uniform distribution (the 'slab'). The first element of the mixture (the 'spike') assigns all its probability to zero, and the Bernouilli element gives the prior probability of that not being appropriate for a single regression coefficient. As such, the Bernoulli probability may be termed the probability of selection. The second element of the mixture (the 'slab') is used as an uninformative prior within the bounds (a,b), reflecting that this is an exploratory tool and the user has no prior knowledge of the effect that could be expected for the predictors to be considered. The mixture distribution is represented graphically in Figure 1. For a more formal - yet accessible - explanation of the working of spike-and-slab and other Bayesian selection models see Forte et al. (2018).
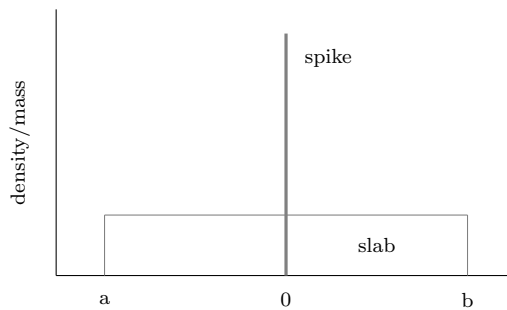


**Fig. 1** Spike and slab prior distribution

In our analysis we use the 'BoomSpikeSlab' package (Scott, 2022) in R, and different sets of priors for the probability of selection. We start by estimating a standard binary logit model using frequentist statistics, including all the predictors listed in Table 1. We call this Model 1, which can be taken as the approach that would be followed if problems related to model overfitting are neglected. We would expect this to be the case if theory, past research, or current policy are used as the criteria to select predictors. Based on this standard modelling approach, all available predictors would be included since they are either theoretically pertinent, listed in the sentencing guidelines as relevant, or, like gender, have been found to have a strong effect in the probability of receiving a custodial sentence in the literature (Pina-Sánchez and Harris, 2020).

Model 1 is therefore used as the benchmark to assess the effectiveness of selection models based on Bayesian selection. We estimate two of them, Model 2 and 3, each with a different set of prior probabilities of selection, 0.6 and 0.1, respectively. A prior selection probability of 0.6 approximately returns the twenty most relevant predictors (i.e. their posterior probability of selection will be meaningfully different from zero). This is calculated by taking the ratio of the target number of predictors that we would like to keep in our model over the total of candidate predictors considered; in our case, $20/34 = 0.59$. The inclusion prior probability of 0.1 is used to explore an extreme selection process, where only those factors that can be identified as the most important case characteristics are selected, i.e. those for which the effect is so strong that even a small prior probability of inclusion will not be enough to drop them from the model. In all our spike-and-slab models we take predictors with posterior selection probabilities smaller than 0.05 as effectively dropped from the model.

## 5 Results: Factors Predicting Magistrates' Court Sentencing of Shoplifting Offences

Results from the four models of interest are reported in Table 2. The effect estimated for each case characteristics is reported using odds ratios. Hence, case characteristic with an odds ratio below one reduce the probability of receiving a custodial sentence, and those with odds ratios above one increase it. Measures of uncertainty for the odds ratios are expressed using 95% confidence and credible intervals, depending on whether the models are based on frequentist or Bayesian statistics.

Model 1, where all predictors are included, shows seventeen statistically significant predictors, namely half of the set of 34 included in the model. Such a relatively low proportion of case characteristics found to be significantly predicting sentence severity can be explained by their relatively low prevalence (sixteen case characteristics were present in less than 10% cases from our sample), but also by the expected problem of model overfit. Specifically, the multiple non-significant predictors left in the model unnecessarily reduce its degrees of freedom, and can lead to multicollinearity. Multicollinearity is in fact found to be present in the aggravating factor previous convictions ($VIF = 3.70$), which appears to have an extreme effect (19.81 odds ratio) that is highly unreliable (with a 95% confidence interval ranging from 6.85 to 85.57 odds ratio).

Model 2, the first of our spike and slab models where the prior probability of inclusion is set at 0.6, drops fifteen of the predictors present in Model 1 (i.e. their estimated posterior probability being different from zero is lower than 0.05), but in doing so it becomes a better model. This is because the smaller set of predictors reduces overfitting; and that is achieved while increasing its classification accuracy. From a 58% correct classification obtained by the null model (i.e. classifying all sentences as the modal category, non-custody), Model 1 classifies 80.7% of sentence outcomes accurately, whereas Model 2 reaches an 81.2% accuracy rate. Similarly, we can also see that previous issues of multicollinearity have been resolved, with the estimate for the aggravating factor previous convictions halving its effect size (9.56 odds ratio) and its 95% credible interval becoming narrower (4.38 to 20.84 odds ratio).

In short, Model 2 (based on Bayesian selection) is a more parsimonious model, improving Model 1 (based on a standard frequentist approach) in multiple ways. Namely, Model 2 offers a more robust capacity to detect factors affecting custodial decisions, more accurate and precise estimates of the effects associated to those factors, and higher classification accuracy. Furthermore, the selection process was undertaken by considering all predictors simultaneously, which eliminates arbitrary selection choices (forward, backward, etc.) that can affect the final set of selected predictors. This is another key feature of the spike and slab approach presented here.

Model 3 shows how the list of 'most important factors' can be further trimmed simply by reducing the prior probability of inclusion. If we do so in a rather extreme way, namely by reducing the prior inclusion probability from 0.6 to 0.1, we still find thirteen case characteristics are selected by the model. This is a surprisingly large number of factors predicting magistrates' court sentencing, and particularly given the low prior probability of selection (0.1), the relatively small sample size ($N = 2,116$), and the fact that the sample is comprised of a single, highly homogenous, offence type (shoplifting). This finding suggests that the sentencing guidelines in England and Wales have not led to uniform sentencing through the stifling of judicial discretion (at least for shoplifting offences) as in some other jurisdictions (Dhami et al., 2015), but rather they facilitate the distinction of cases according to the relevant features that define them (Marder and Pina-Sánchez, 2020; Roberts et al., 2018).

By reviewing the set of factors selected in Model 3 we can obtain a more robust understanding of the sentencing of shoplifting offences in the magistrates' courts in England and Wales. The two case characteristics with the largest effect size are previous convictions and caring responsibilities. Their effect is so strong (9.48 and 0.05 odds ratios, respectively) that in many cases the presence of these factors alone comes close to determining whether the offender will receive a custodial sentence. For example, if we consider a reference case defined by, say, an offender found to use professional methods to steal £500 to £1000 worth of goods, with no other case characteristics deemed relevant, the estimated

probability of receiving a custodial sentence is 0.14, but the probability for the same case goes up to 0.61 if relevant previous convictions are deemed to be present, and down to 0.01 if instead caring responsibilities are present. The large effect of previous conviction in cases of theft has been previously identified in the sentencing literature (Sentencing Council, 2015c, 2019; Roberts and Pina-Sánchez, 2014), however, the even larger effect seen for caring responsibilities contradicts findings reported from more serious offences sentenced in the Crown Court (Kane and Minson, 2022).

**Table 2** Results for the logit models specifying the probability of receiving either a suspended or immediate custodial sentence for a sample of offenders of shoplifting sentenced in the Magistrates' Court

| | Model 1: Freq. | | Model 2: Prior P= 0.6 | | | Model 3: Prior P= 0.1 | | |
|---|---|---|---|---|---|---|---|---|
| | OR | 95% CI | OR | 95% CI | Post. P | OR | 95% CI | Post. P |
| Intercept | 0.01 | (0.01, 0.03) | 0.02 | (0.01, 0.06) | 1.00 | 0.02 | (0.01, 0.07) | 1.00 |
| Culpability: level of planning | 1.32 | (1.04, 1.67) | 1.38 | (0.93, 2.05) | 0.76 | 1.04 | (0.82, 1.32) | 0.12 |
| Culpability: use of force | 1.20 | (0.75, 1.92) | | | | | | |
| Culpability: role of offender | 1.11 | (0.93, 1.31) | 1.02 | (0.91, 1.13) | 0.16 | | | |
| Culpability: sophisticated offence | 3.07 | (1.27, 7.60) | 1.85 | (0.32, 10.67) | 0.35 | 2.20 | (0.38, 12.75) | 0.46 |
| Culpability: banning order | 0.83 | (0.31, 2.18) | | | | | | |
| Culpability: involvement of others | 0.74 | (0.27, 1.90) | | | | | | |
| Culpability: mental disorder | 0.30 | (0.13, 0.66) | 0.29 | (0.12, 0.70) | 0.95 | 0.34 | (0.08, 1.48) | 0.70 |
| Harm: value of goods stolen | 1.33 | (1.22, 1.46) | 1.34 | (1.16, 1.54) | 0.97 | 1.39 | (1.22, 1.59) | 0.98 |
| Harm: emotional distress | 1.33 | (0.50, 3.45) | | | | | | |
| Harm: injury to victim | 5.61 | (1.73, 20.11) | | | | | | |
| Harm: effect on business | 0.94 | (0.73, 1.20) | | | | | | |
| Harm: other factors | 0.90 | (0.54, 1.47) | | | | | | |
| Aggravating: previous convictions | 19.81 | (6.85, 85.57) | 9.56 | (4.38, 20.84) | 0.99 | 9.48 | (3.45, 24.90) | 0.99 |
| Aggravating: conceal evidence | 1.20 | (0.76, 1.86) | | | | | | |
| Aggravating: failure to comply | 4.41 | (3.38, 5.76) | 4.45 | (2.95, 6.71) | 0.99 | 4.19 | (2.80, 6.26) | 0.99 |
| Aggravating: offender on bail | 5.32 | (3.97, 7.16) | 5.30 | (3.23, 8.71) | 0.99 | 5.06 | (3.11, 8.26) | 0.99 |
| Aggravating: offences into consideration | 1.32 | (0.93, 1.86) | 1.04 | (0.82, 1.33) | 0.12 | | | |
| Aggravating: harm to the community | 7.22 | (2.45, 24.27) | 3.57 | (0.53, 24.12) | 0.71 | 1.89 | (0.28, 12.55) | 0.33 |
| Aggravating: professional offending | 2.87 | (1.66, 5.00) | 3.59 | (1.68, 7.65) | 0.99 | 4.33 | (2.23, 8.43) | 0.99 |
| Aggravating: stealing goods to order | 1.58 | (0.79, 3.16) | 1.03 | (0.81, 10.67) | 0.09 | | | |
| Aggravating: other factors | 2.31 | (1.77, 3.03) | 2.38 | (1.65, 3.38) | 0.99 | 2.42 | (1.62, 3.48) | 0.98 |
| Mitigating: lack of maturity | 0.50 | (0.21, 1.10) | 0.93 | (0.58, 1.48) | 0.12 | | | |
| Mitigating: good character | 1.24 | (0.25, 4.62) | | | | | | |
| Mitigating: no recent convictions | 3.41 | (0.89, 16.95) | | | | | | |
| Mitigating: financial hardship | 0.62 | (0.39, 0.97) | 0.90 | (0.59, 1.37) | 0.26 | | | |
| Mitigating: steps to address addiction | 0.48 | (0.33, 0.70) | 0.56 | (0.28, 1.10) | 0.79 | 0.90 | (0.58, 1.39) | 0.20 |
| Mitigating: mental disorder | 0.71 | (0.40, 1.21) | | | | | | |
| Mitigating: remorse | 0.56 | (0.37, 0.81) | 0.49 | (0.33, 0.74) | 0.99 | 0.60 | (0.26, 1.35) | 0.65 |
| Mitigating: return of stolen property | 0.34 | (0.10, 0.94) | | | | | | |
| Mitigating: serious medical condition | 1.29 | (0.62, 2.59) | | | | | | |
| Mitigating: primary carer | 0.11 | (0.01, 0.51) | 0.13 | (0.02, 0.92) | 0.85 | 0.05 | (0.01, 0.59) | 0.99 |
| Mitigating: other factors | 0.56 | (0.38, 0.80) | 0.62 | (0.34, 1.12) | 0.79 | | | |
| Offender: age band | 1.00 | (0.88, 1,14) | | | | | | |
| Offender: male | 1.43 | (1.08, 1.90) | | | | | | |
| Percentage correctly classified | 80.7% | | 81.2% | | | 80.1% | | |
| $N = 2,116$ | | | | | | | | |

Note 1: 'Freq.' stands for frequentist, 'Prior P' for prior probability of inclusion, 'Post. P' for posterior probability of inclusion, 'OR' for odds ratios and 'CI' for confidence/credible interval.

Note 2: Results for Model 2 and 3 reflect the posterior summaries of the coefficients conditional on being different from zero.

More broadly, these findings question the view that sentencing in England and Wales is based on the principle of proportionality. Studies exploring samples composed of different offence types have found the expected relationship between sentence severity and offence seriousness. However, when we focus on a single offence type, as we do here, we note how offender related factors (such as the number of previous convictions, or whether a primary carer) appear to be far more influential in determining sentence severity than factors defining the offence (such as the value of the goods stolen or its lasting effect on the victim).

Further comparisons across groups of sentencing guideline factors also show an unexpected higher relevance of Step Two factors (aggravating and mitigating factors) over the a priori more consequential (Mitchell, 2013) Step One factors (factors deemed to be relevant to judging harm and culpability). Step One factors are meant to provide the main factual elements of the offence, determining the starting point for the sentencing decision, while Step Two provide context and are meant to 'fine-tune' the starting point determined at Step One (Ashworth and Roberts, 2013). However, in Model 3, nine Step Two factors were found to predict custodial sentences compared to just four Step One factors. Thus, the factual elements defining the offence appear to be less important than the more general factors used to contextualise that offence - a result that was also detected in the impact assessment of the guideline (Sentencing Council, 2019), and in Pina-Sánchez et al. (2020) analysis of sentencing of assault cases in the Crown Court. Future sentencing guidelines could redress this problem by expanding the range of starting points following Step One considerations of offence seriousness.

We also observe a relatively stronger effect of aggravating factors (with six of them shown to predict sentence severity) compared to mitigating factors (with only three selected by the model). This is an imbalance anticipated by Padfield (2013) and Hutton (2013), who pointed at the stronger emphasis and larger number of aggravating factors, compared to mitigating factors, listed in the guidelines. It is also worth noting that no demographic disparities are detected since neither age nor gender were selected by the model; and offenders' ethnicity was not recorded in the dataset. These results contrast with the gender disparity observed by Pina-Sánchez and Harris (2020) in the sentencing of assault, drugs and burglary offences by the Crown Court, where the odds ratio of receiving a custodial sentence for male offenders was approximately 200% to 300% higher than for female offenders.

## 6 Discussion: A New Analytical Strategy, Empirical Insights, and Avenues for Sentencing Research

Understanding sentencing decisions often has direct real-world applications. For example, informing sentencers about current sentencing practices helps promote greater consistency, as decisions are reconsidered in the light of practices followed in other courts outside their direct contact (Hester, 2017; Pina-Sánchez et al., 2019). Similarly, attempts at sentencing reform, such as calls for reducing the use of custodial sentences for female offenders (Corston, 2007), fall short of their potential reach when the largely inconsequential effect of heavily gendered mitigating factors, such as mental health condition, remains unacknowledged (Justice Committee, 2022).

Here, we have contributed to the refinement and diversification of the empirical toolbox for sentencing research by showcasing the potential of spike-and-slab models, a type of Bayesian model selection technique that has not yet been used in the study of judicial decision-making. This technique shows particular promise for the analysis of sentencing datasets where the number of cases is small relative to the number of case characteristics to be explored. In these circumstances, standard regression models analysing sentencing decisions can lead to wrong inferences as they are easily overfitted when all case characteristics are introduced in the model, or prone to arbitrary decisions when selection models are used, or when case characteristics are not introduced in the model without a priori clear justification.

We estimated spike-and-slab models using a new dataset of shoplifting offenders sentenced in the magistrates' courts in England and Wales. In spite of the limited sample size, the spike-and-slab models

allowed us to provide unbiased selections of the most influential predictors of custodial sentences from the initial set of 34. In contrast to an approach where all case characteristics are considered, we noted how our selection models discarded the least influential factors, reducing multicollinearity, and providing more accurate estimates and measures of uncertainty for those case characteristics retained by the model.

From an empirical perspective, we have drawn a number of useful insights into the sentencing of shoplifting offenders in the magistrates' court. These include the finding that a relatively large number of case characteristics predict sentence severity (i.e., custodial sentences); that Step Two factors appear to be more important than Step One factors; and that factors defining the offender rather than the offence have the greatest effects. We have also seen how the set of most relevant case characteristics predicts the final sentence with a fair degree of accuracy (over 80% classification accuracy), which corroborates similar predictive rates seen in the Crown Court (Pina-Sánchez and Grech, 2017). This predictive rate points at judicial decisions being fairly predictable, but not deterministic, even when a wide range of case characteristics defining the case are known, which lends support to the view that sentencing in England and Wales is neither a science nor an art (Hooper, 2015).

By comparing the effects estimated for different case characteristics listed in the guidelines we have also obtained some useful insights. We find that Step Two factors appear to be more influential than Step One factors, which contradicts the expected functioning of the guidelines. This is the case since Step One factors are meant to provide the main factual elements of the offence, determining the starting point for the sentencing decision, while Step Two provide context and are meant to 'fine-tune' the starting point determined at Step One (Ashworth and Roberts, 2013). Perhaps more importantly, this finding questions the widely held belief that sentencing in England and Wales is centered around the principle of proportionality. Studies exploring samples composed of different offence types have found the expected relationship between sentence severity and offence seriousness. However, when we focus on a single offence type, as we do here, we note how offender related factors (such as the number of previous convictions, or whether a primary carer) appear to be far more influential in determining sentence severity than factors defining the offence (such as the value of the goods stolen or its lasting effect on the victim).

Besides identifying the most influential case characteristics, spike-and-slab models could also be applied to a wide range of recurrent sentencing research questions. For example, the unique exploratory capacity of such models can be employed to identify potential interactions in the way case characteristics are applied by sentencers, which is a growing area of research in the sentencing literature (Belton, 2018; Maslen, 2015; Lightowlers and Pina-Sánchez, 2018; Lightowlers, 2019). There is a consensus in the sentencing literature that case characteristics are not considered in isolation, but rather their effect size is contingent on the presence of other characteristics featuring in the same case (Jacobson and Hough, 2007). However, due to the large sample sizes required for the estimation of interaction effects (Greenland, 1983), these studies have often been restricted to the testing of one or a few interactions. The model selection capabilities of spike-and-slab specifications eliminate this restriction and potentially allow researchers to test for all two-way combinations of case characteristics contained in the sentencing guidelines in a principled and computationally efficient way.

Lastly, we believe it would be particularly important to undertake a similar exploratory analysis at scale with the goal of detecting factors conducive of unwarranted influences. So far, the large literature exploring biases and disparities in sentencing has focused on specific themes such as the socio-demographic characteristics of the offender (King and Johnson, 2016), judge (Drápal and Pina-Sánchez, 2022), court (Pina-Sánchez and Grech, 2017), and geographic location (Fearn, 2005), or even more apparently spurious factors such as the weather (Drápal and Pina-Sánchez, 2019), and sport results (Eren and Mocan, 2018). However, it is possible that many of those extraneous factors are confounding each other, and that many other factors conducive of unwarranted disparities remain unknown. Hence, the need for a wider, more comprehensive, exploratory analysis, where a larger number

of predictors is considered, potentially leading to similar modelling challenges as those faced here, which could again be resolved using spike-and-slab, or similar Bayesian selection models.

# References

Ashworth, A. and J. V. Roberts (2013). The origins and nature of the sentencing guidelines in England and Wales. In A. Ashworth and J. V. Roberts (Eds.), *Sentencing Guidelines: Exploring the English Model*, Oxford, pp. 1–12. Oxford University Press.

Belton, I. (2018). *The role of personal mitigating factors in criminal sentencing judgments: An empirical investigation.* Ph. D. thesis, Middlesex University.

Corston, J. (2007). The corston report: A review of women with particular vulnerabilities in the criminal justice system. Technical report, Home Office.

Dhami, M. K. (2013). A 'decision science' perspective on the old and new sentencing guidelines in England and Wales. In A. Ashworth and J. V. Roberts (Eds.), *Sentencing Guidelines: Exploring the English Model*, Oxford, pp. 165–181. Oxford University Press.

Dhami, M. K. (2022). Sentencing multiple-versus single-offence cases: Does more crime mean less punishment? *The British Journal of Criminology 62*(1), 55–72.

Dhami, M. K. and I. Belton (2015). Using court records for sentencing research: Pitfalls and possibilities. In J. V. Roberts (Ed.), *Exploring sentencing practice in England and Wales*, Oxford, pp. 18–34. Palgrave Macmillan.

Dhami, M. K., I. Belton, and J. Goodman-Delahunty (2015). Quasi-rational models of sentencing. *Journal of Applied Research on Memory and Cognition 4*, 239–247.

Dhami, M. K., R. Hertwig, and U. Hoffrage (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin 6*(130), 959–988.

Drápal, J. and J. Pina-Sánchez (2019). Does the weather influence sentencing? empirical evidence from czech data. *International Journal of Law, Crime and Justice 56*, 1–12.

Drápal, J. and J. Pina-Sánchez (2022). What is the value of judicial experience? exploring judge trajectories using longitudinal data. *Justice Quarterly*, 1–30.

Eren, O. and N. Mocan (2018). Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics 10*(3), 171–205.

Fearn, N. E. (2005). A multilevel analysis of community effects on criminal sentencing. *Justice Quarterly 22*(4), 452–487.

Fleetwood, J., P. Radcliffe, and A. Stevens (2015). Shorter sentences for drug mules: The early impact of the sentencing guidelines in England and Wales. *Drugs: Education, Prevention and Policy 22*(5), 428–436.

Forte, A., G. Garcia-Donato, and M. Steel (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review 86*(2), 237–258.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power. *Statistics in medicine 2*(2), 243–251.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* New York: Springer.

Hester, R. (2017). Judicial rotation as centripetal force: Sentencing in the court communities of South Carolina. *Criminology 55*(1), 205–235.

Hooper, A. (2015). Sentencing: art or science. *Singapore Academy of Law Journal 27*(1), 17–30.

Hutton, N. (2013). The definitive guideline on assault offences: The performance of justice. In A. Ashworth and J. Roberts (Eds.), *Sentencing Guidelines: Exploring the English Model*, Oxford, pp. 86–103. Oxford University Press.

Isaac, A., J. Pina-Sánchez, and A. Varela (2021). The impact of three guidelines on consistency in sentencing. *Sentencing Council for England and Wales*.

Jacobson, J. and M. Hough (2007). Mitigation: The role of personal factors in sentencing. Technical report, Prison Reform Trust.

Justice Committee (2022). Women in prison. Technical report.

Kane, E. and S. Minson (2022). Analysing the impact of being a sole or primary carer for dependent relatives on the sentencing of women in the Crown Court, England and Wales. *Criminology and Criminal Justice*.

King, R. D. and B. D. Johnson (2016). American journal of sociology. *A punishing look: Skin tone and Afrocentric features in the halls of justice 122*(1), 90–124.

Langan, P. A. and D. J. Levin (2002). *Recidivism of prisoners released in 1994*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Lightowlers, C. (2019). Drunk and doubly deviant? the role of gender and intoxication in sentencing assault offences. *The British Journal of Criminology 59*(3), 693–717.

Lightowlers, C. and J. Pina-Sánchez (2018). Intoxication and assault: An analysis of Crown Court sentencing practices in England and Wales. *The British Journal of Criminology 58*(1), 132–154.

Marder, I. and J. Pina-Sánchez (2020). Nudge the judge? Theorising the interaction between heuristics, sentencing guidelines and sentence clustering. *Criminology & Criminal Justice 20*(4), 399–415.

Maslen, H. (2015). Penitence and persistence: How should sentencing factors interact? In J. Roberts (Ed.), *Exploring Sentencing Practice in England and Wales*, pp. 173–193. Basingstoke: Palgrave Macmillan.

Mitchell, B. (2013). Sentencing guidelines for murder: From political schedule to principled guidelines. In A. Ashworth and J. V. Roberts (Eds.), *Sentencing Guidelines: Exploring the English Model*, Oxford, pp. 52–70. Oxford University Press.

Mitchell, T. J. and J. J. Beauchamp (1998). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Padfield, N. (2013). Exploring the success of sentencing guidelines. In *Sentencing Guidelines: Exploring the English Model*, pp. 31–51. Oxford: Oxford University Press.

Pina-Sánchez, J., I. Brunton-Smith, and L. Guangquan (2020). Mind the step: A more insightful and robust analysis of the sentencing process in England and Wales under the new sentencing guidelines. *Criminology & Criminal Justice 20*(3), 268–301.

Pina-Sánchez, J. and J. P. Gosling (2020). Tackling selection bias in sentencing data analysis: a new approach based on a scale of severity. *Quality & Quantity 54*, 1047–1073.

Pina-Sánchez, J., J. P. Gosling, H. Chung, S. Geneletti, E. Bourgeois, and I. Marder (2019). Have the England and Wales guidelines influenced sentence severity? An empirical analysis using a scale of sentence severity and time-series analyses. *British Journal of Criminology 59*(4), 979–1001.

Pina-Sánchez, J. and D. Grech (2017). Location and sentencing: To what extent do contextual factors explain between court disparities? *British Journal of Criminology 58*(3), 529–549.

Pina-Sánchez, J., D. Grech, I. Brunton-Smith, and D. Sferopoulos (2019). Exploring the origin of sentencing disparities in the Crown Court: Using text mining techniques to differentiate between court and judge disparities. *Social Science Research 84*, 1–13.

Pina-Sánchez, J. and L. Harris (2020). Sentencing gender? investigating the extent and origin of sentencing gender disparities in the Crown Court. *Criminal Law Review 1*, 3–28.

Roberts, J. V. (2013). Complying with sentencing guidelineslatest findings from the Crown Court Sentencing Survey. In *Sentencing Guidelines: Exploring the English Model*, pp. 104–121. Oxford: Oxford University Press.

Roberts, J. V. and J. Pina-Sánchez (2014). The role of previous convictions at sentencing in the Crown Court: Some new answers to an old question. *Criminal Law Review 8*, 575–588.

Roberts, J. V., J. Pina-Sánchez, and I. Marder (2018). Individualisation at sentencing: The effects of guidelines and 'preferred' numbers. *Criminal Law Review 2*, 123–136.

Ročková, V. and E. I. George (2014). Negotiating multicollinearity with spike-and-slab priors. *Metron 72*(2), 217–229.

Scott, S. L. (2022). Boomspikeslab: Mcmc for spike and slab regression. Technical report, The Comprehensive R Archive Network.

Sentencing Council (2014). Theft offences: Analysis and research bulletin. Technical report.

Sentencing Council (2015a). Annex b: Quality and methodology note. Technical report.

Sentencing Council (2015b). Theft offences: Definitive guideline. Technical report.

Sentencing Council (2015c). Theft offences: Research report. Technical report.

Sentencing Council (2019). Theft offences: Assessment of guideline. Technical report.

Smith, G. (2018). Step away from stepwise. *Journal of Big Data 5*(1), 1–12.

Sturge, G. (2021). Court statistics for England and Wales. Technical report, House of Commons Library.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.