Smart Data Insights into Online Gambling: A Geodemographic Analysis of Behavioural Clusters in Great Britain

Shunya Kimura*[1], Justin van Dijk†[1] and Paul Longley‡[1]

[1]University College London, Gower Street, WC1E 6BT London, UK

GISRUK 2025

Summary

Online gambling is a rapidly growing pastime, offering convenience and entertainment. However, concerns persist regarding its potential harms and how individuals engage with online platforms. This thesis contends that survey-based investigations often yield incomplete representations of gambling behaviours, hindered by self-report biases and small sample sizes. To address this, a data-intensive analysis of ~1.2 million online accounts from a major British operator characterises gambling behaviours in Great Britain throughout 2022. Using clustering techniques, the study identifies 12 distinct behavioural typologies. Findings underscore the value of Smart Data in shaping policy, informing interventions and advancing research on gambling's societal impact.

KEYWORDS: geodemographics, online gambling, public health, Smart Data

1 Introduction

Within Great Britain (GB), a surge in gambling behaviour studies over the past decade underscores the importance of addressing associated harms as a potential public health issue (Pickering & Blaszczynski, 2021; Wardle et al., 2024). Yet, significant knowledge gaps remain — particularly in relation to online gambling — exacerbated by limitations in existing evidence. These include measurement errors in surveys, such as social desirability bias and recall inaccuracies, as well as restricted sampling fractions due to the high costs of data collection and the relatively low prevalence of gambling disorder in the population (Sturgis & Kuha, 2021). The current United Kingdom (UK) initiatives in *Smart Data* (Department for Business, Energy and Industrial Strategy, 2021) highlights the benefits of cross-sector data sharing in regulated markets, including gambling, to enhance innovation and consumer outcomes. Studies adopting Smart Data from industry partners have advanced this agenda, uncovering revealed behaviours that would likely remain overlooked or inaccessible through conventional methods (Broda et al., 2008; Forrest & McHale, 2022; Rains &

*shunya.kimura.18@ucl.ac.uk
†j.vandijk@ucl.ac.uk
‡p.longley@ucl.ac.uk

Longley, 2021). Crucially, such data stem from consumer records offering far greater detail and geographic granularity than conventional survey-based approaches, unencumbered by non-response or recall errors, and thus providing more direct and reliable indicators of gambling behaviours. Building on these developments, this study analyses 12-months of data from one of Britain's leading gambling operators, focusing on geodemographic segmentation to reveal socio-spatial dimensions of gambling activity and its potential public health implications.

2 Methods

2.1 Data Description

This study draws on proprietary data from one of the 'Big Five' British gambling operators, for which we were granted an independent, time-limited access to their extensive data warehouse. It is important to underscore that the operator exerts no influence over the research design. In light of broader concerns within the research community about potential industry influence (Wardle et al., 2024), this study adopts a reflexive and transparent approach throughout. We focus on a 12-month period spanning 2022, examining both behavioural records and account registration information for 1,184,905 *Genuine*[1] Customers, consisting of both *Active*[2] and *Dormant*[3] Gamblers.

Behavioural records include both gambling activities and monetary transactions related to deposits and withdrawal: often referred to as 'account behaviour' (PwC, 2017) or 'payments transactions' (Ghaharian et al., 2023). Transaction-level records for each individual are systematically processed and aggregated to derive 37 behavioural features, representing average annual gambling activity (see Table 1). These features are organised into four conceptual domains, extending Braverman and Shaffer (2010)'s framework: *frequency*, *intensity*, *riskiness* and *variability* of play. Active Gamblers are then grouped according to their mode of gambling involvement:

- **Group BG: Betting-and-Gaming**: individuals who participate in both betting[4] and gaming[5].
- **Group B: Betting-Exclusive**: individuals who place bets but do not engage in gaming.
- **Group G: Gaming-Exclusive**: individuals who gamble only through gaming products and never place bets.

Account registration information include age, gender and home address. Collectively, these data form the foundation of this study.

[1] *Genuine Customers* are defined as individuals registered with a valid GB address, engaging in at least one form of gambling using money deposits (rather than free or demo modes) and making deposits on multiple days within the year.

[2] *Active Gamblers* are defined as Genuine Gamblers who meet the criteria for sustained gambling activity, including gambling at least once a month on average and maintaining an account tenure of over three months.

[3] *Dormant Gamblers* are Genuine Gamblers who do not meet the additional criteria for Active Gamblers.

[4] *Betting* refers to gambling activities recorded through discrete bet slips as with, for example, football, horse racing and combat sports bets.

[5] *Gaming* refers to gambling activities recorded as continuous sessions, including slots, jackpot slots and bingo.

Table 1

Domain / Feature Name	Description
Frequency	
Tenure (in days)	The total number of days a customer has been registered.
Avg. gap between deposit-days	The average number of days between wagers.
Avg. gap between gambling-days	The average number of days between deposits.
Withdrawal-deposit ratio	The ratio of total number of withdrawal days to deposit days.
Gambling-deposit ratio	The ratio of total number of gambling days to deposit days.
Prop. of gambling-active days (%)	The percentage of registered days a customer engaged in gambling.
Intensity	
Avg. monthly deposit amount (£)	The average total deposit amount per month.
Avg. monthly deposit quantity	The average total number of deposits per month.
Avg. deposit amount per deposit-days (£)	The average deposit amount on days when deposits are made.
Avg. stake amount per gambling-days (£)	The average amount wagered on days when gambling occurs.
Avg. monthly loss amount (£)	The average monthly net loss, calculated as total losses divided by months registered.
Avg. stake amount per bet slip (£)	The average amount wagered per bet slip.
Avg. betting-day bet count	The average number of bets placed on betting days.
Avg. stake amount per session (£)	The average amount wagered per gaming session.
Avg. gaming-day session count	The average number of gaming sessions per gaming day.
Avg. session interaction	The average level of engagement per gaming session.
Avg. session duration (sec.)	The average duration per gaming session.
Riskiness	
Monthly loss-deposit ratio	The average proportion of deposited funds lost each month.
Prop. of loss-days (%)	The percentage of gambling days where losses occurred.
Avg. potential return per bet (£)	The average expected payout per bet slip, accounting for wagered amounts and odds.
Avg. fold quantity per bet	The average number of individual bets (folds) within each bet slip.
Prop. of acca bets (%)	The percentage of bets that are accumulator (acca) bets.
Variability	
Std. deviation gap between deposit-days	The variation in time intervals between deposit days.
Std. deviation gap between gambling-days	The variation in time intervals between gambling days.
Std. deviation stake amount per gambling-days (£)	The variation in stake amounts on gambling days.
Prop. of popular-day plays	The percentage of gambling activity occurring on the 92 most popular gambling days.
Prop. of weekend plays	The percentage of gambling activity occurring on weekends.
Std. deviation stake amount per bet slip	The variation in stake amounts per bet slip.
Std. deviation betting-day bet count	The variation in the number of bets placed on betting days.
Std. deviation potential return per bet	The variation in expected payouts per bet slip.
Prop. of late-night bets	The percentage of bets placed between midnight and 5:59 am.
Total no. of activities bet	The total number of different betting activities a player participated in.
Std. deviation stake amount per session	The variation in stake amounts per gaming session.
Std. deviation gaming-day session count	The variation in the number of gaming sessions per day.
Std. deviation session duration (sec.)	The variation in gaming session lengths.
Prop. of late-night sessions (%)	The percentage of gaming sessions occurring between midnight and 5:59 am.
Total no. of activities gamed	The total number of different gaming activities a player participated in.

Variables used in analysis.

2.2 A Stepwise Framework for Active Gambler Segmentation

Our objective is to segment 796,378 Active Gamblers into distinct clusters (or Subgroups) based solely on revealed gambling behaviours. Prior to clustering, input variables are transformed using the Inverse Hyperbolic Sine (IHS) function, given that many features are strongly right-skewed and contain zero or negative values (Bellemare & Wichman, 2020). For certain left-skewed features, namely the *proportion of loss-days*, reflection ($x' = max(x) - x$) preceded IHS transformation to align distributions (Watthanacheewakul, 2021). Transformed data are then normalised to a [0, 100] scale via min-max scaling to ensure that all features are measured on a consistent scale. These pre-processing steps mirror those employed in the 2021/2 and 2011 UK Output Area Classifications (Gale et al., 2016; Wyszomierski et al., 2024).

Following this, a two-step clustering framework is employed separately to the three Active Gambler groups (BG, B and G) — beginning with Principal Component Analysis (PCA) to address collinearity among behavioural features, followed by k-means clustering using the resulting Principal Components (PCs). To determine the initial number of clusters (k) — as required by the algorithm — clustergrams (Fleischmann, 2023) are generated as a visual aid for identifying an appropriate solution. Ultimately, the clustering framework identified six Subgroups for Betting-and-Gaming (BG), two for Betting-Exclusive (B) and three for Gaming-Exclusive (G) (11 total), validated by Total Within-cluster Sum of Squares (TWSS) stability across trials.

Once individuals are assigned to clusters (or Subgroups), they are linked back to their raw behavioural data via their customer IDs, enabling comprehensive *post-hoc* analysis of the identified typologies (see Figure 1 and Figure 2).

2.3 The Socio-Spatial Manifestations of Gambling Behaviours

To assess the relationship between individual gambling patterns and broader neighbourhood-level characteristics, we link customer Subgroups to a range of ancillary data sources from the Geographic Data Service (GeoDS). Specifically, we use the 2021/2 UK Output Area Classification (OAC) and the 2019 harmonised Index of Multiple Deprivation (IMD). The former is the latest residential classification system, utilising small area data from the 2021/2 Census to provide nationwide insights into the broader lifestyle characteristics of neighbourhoods (Wyszomierski et al., 2024). The latter offers a measure of relative social hardship for small areas across GB. To quantify the degree of over- or under-representation of each Subgroup within different neighbourhood types, relative to the national adult population, we apply the Index Score (IS) defined as follows:

$$Index\ Score_{ij} = \frac{\frac{X_{ij}}{\sum X_{ij}}}{\frac{N_i}{\sum N_i}} \times 100 \tag{1}$$

Where:

- X_{ij} is the number of units in category i within group j.
- $\sum X_{ij}$ is the total number of units across all categories within group j.

- N_i is the total size of the reference population in category i.
- $\sum N_i$ is the overall size of the reference population.

The IS compares the observed proportion of customers in a given neighbourhood category to the expected proportion (in this context the adult population of GB); and the Delta Method (Oehlert, 1992) is used to approximate standard errors and construct 95% Confidence Intervals (CIs). An IS below 100 indicates under-representation of gambling behaviour group j in a neighbourhood type i in relative to the GB adult population, whereas an IS above 100 indicates over-representation (see Figure 3 and Figure 4).

2.4 The Geodemographics of Gambling Behaviours

This section outlines the final step in synthesising the findings from individual gambling behaviours and their neighbourhood contexts, thereby making this a geodemographic 'analysis of people by where they live' (Harris et al., 2005). To enable the cluster labelling process, a half-day workshop was held at UCL in collaboration with the gambling operator and data provider. This workshop brings together the operator's senior data analyst and a panel of academic experts, leveraging their domain knowledge to ensure that the labels accurately reflect real-world gambling behaviours. Particular care was taken to avoid value-laden or stigmatising language — such as the term *addict* and *problem gambler* — in line with guidelines for respectful reporting on gambling (Bigger & Wardle, 2024).

3 Results

Figure 1 and Figure 2 present key insights from the behavioural segmentation, illustrating clear distinctions across the Active Gambler cohort. For example, Subgroups BG4 and BG5 exhibit similarly elevated frequencies of play, yet diverge notably in other behavioural dimensions. While BG4 is characterised by low-stake gambling with relatively modest financial exposure, BG5 displays a higher-risk profile marked by high-intensity play — elevated monthly deposit amounts and markedly low win rates, suggesting substantial net losses.

These mix of behaviours are further contextualised by the space in which they occur. The OAC analysis reveals that online gamblers tend to reside in broadly similar types of neighbourhoods, though subtle variations emerge across Subgroups (Figure 3). In general, participants are over-represented in neighbourhoods typified by routine occupations, limited qualifications and higher unemployment — traits aligned with economic precarity. This pattern is reinforced in Figure 4, which shows a strong association between neighbourhood-level deprivation and gambling prevalence. The relationship is particularly pronounced for Group G, while it appears less marked for Group B, suggesting that socio-economic context plays a differential role across gambling modes.

All of these insights — combined with additional behavioural data drawn from the original transactional records, as well as individual demographic information from account registration — are synthesised to assign interpretive labels to the identified clusters (Table 2).

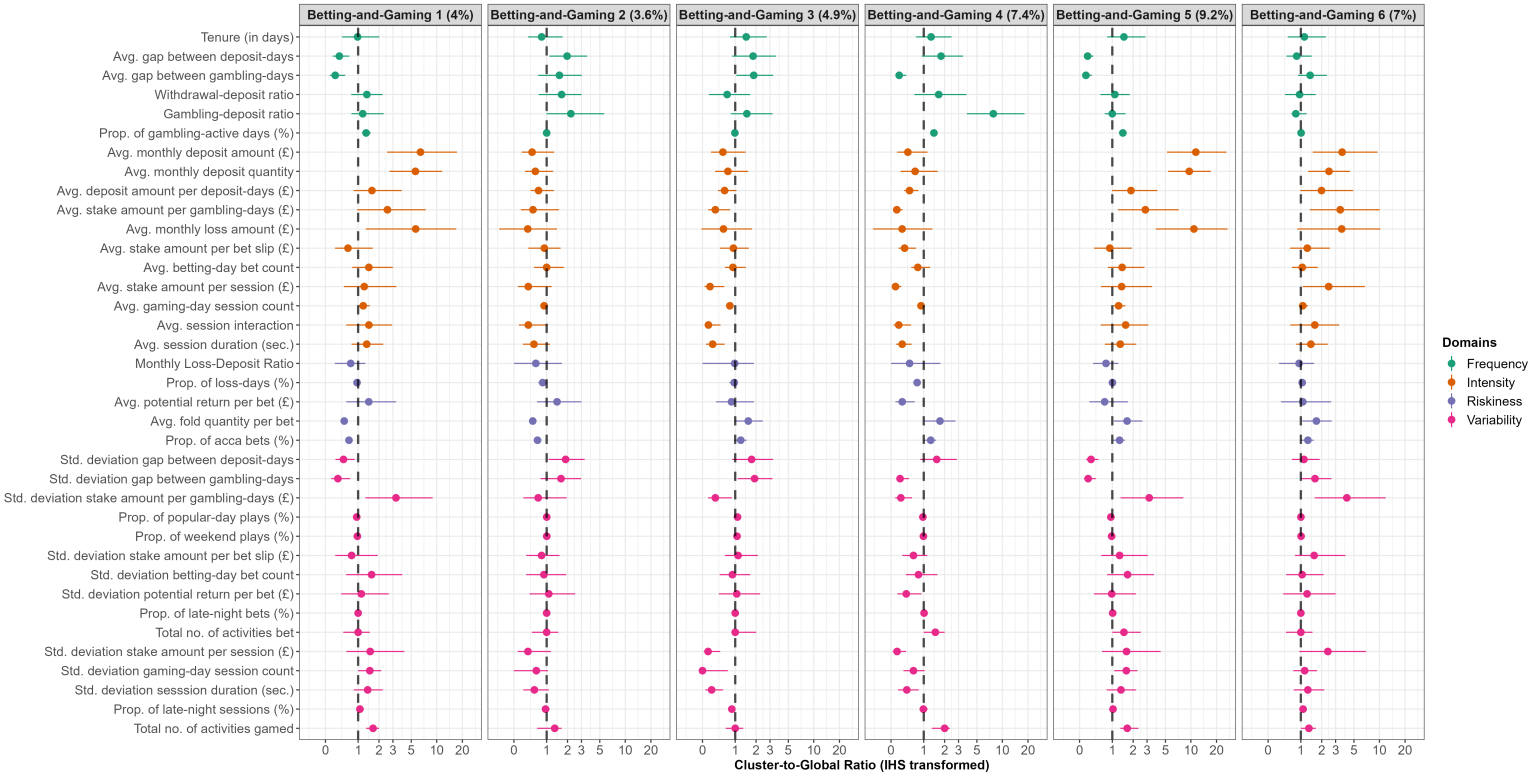

Figure 1: Range plots characterising the six Subgroups within Group BG by 37 features across four domains, relative to the Active Gambler average [the vertical dotted lines indicate the Active Gambler average].

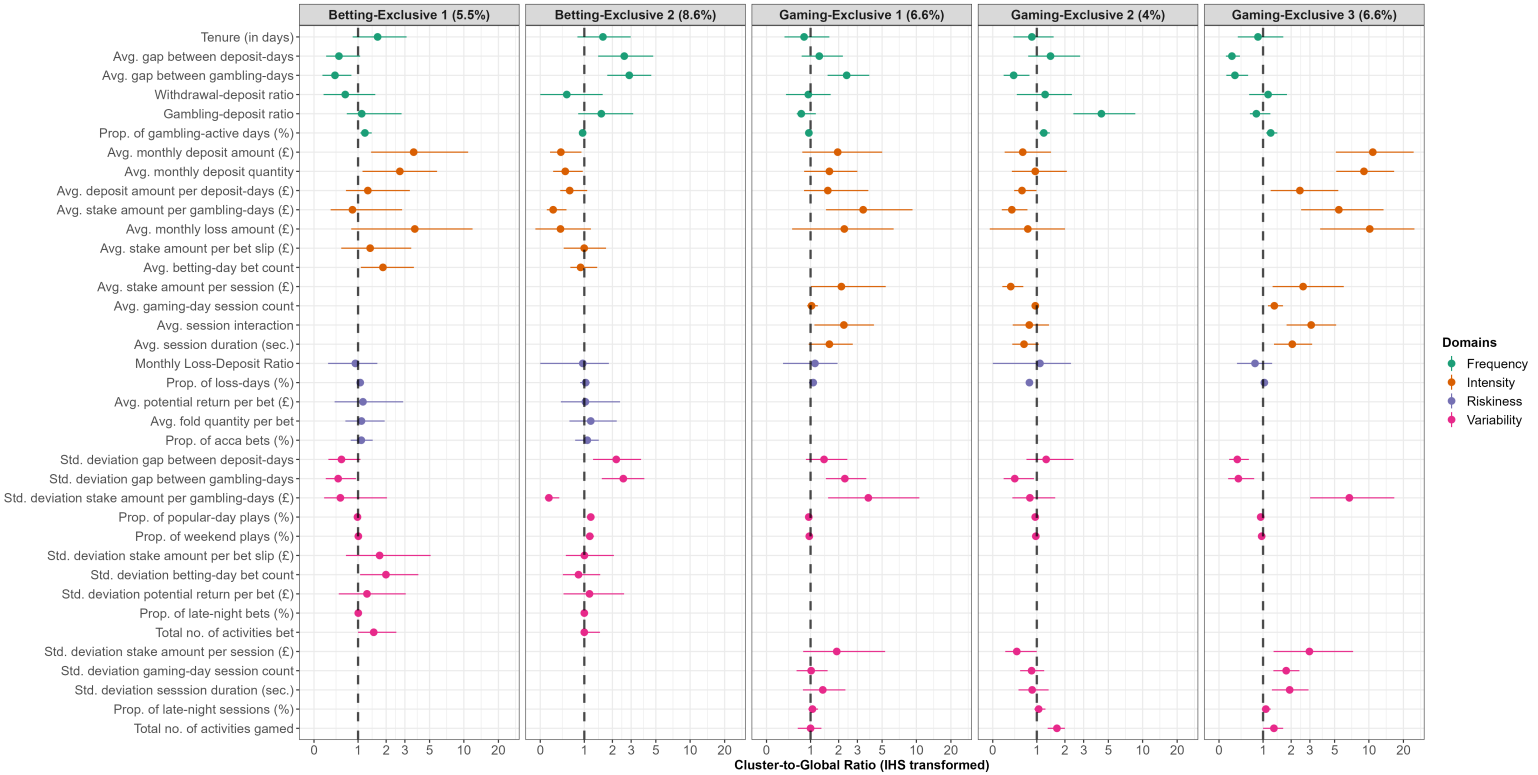

Figure 2: Range plots characterising the five Subgroups within Groups B and G by 37 features across four domains, relative to the Active Gambler average [the vertical dotted lines indicate the Active Gambler average].

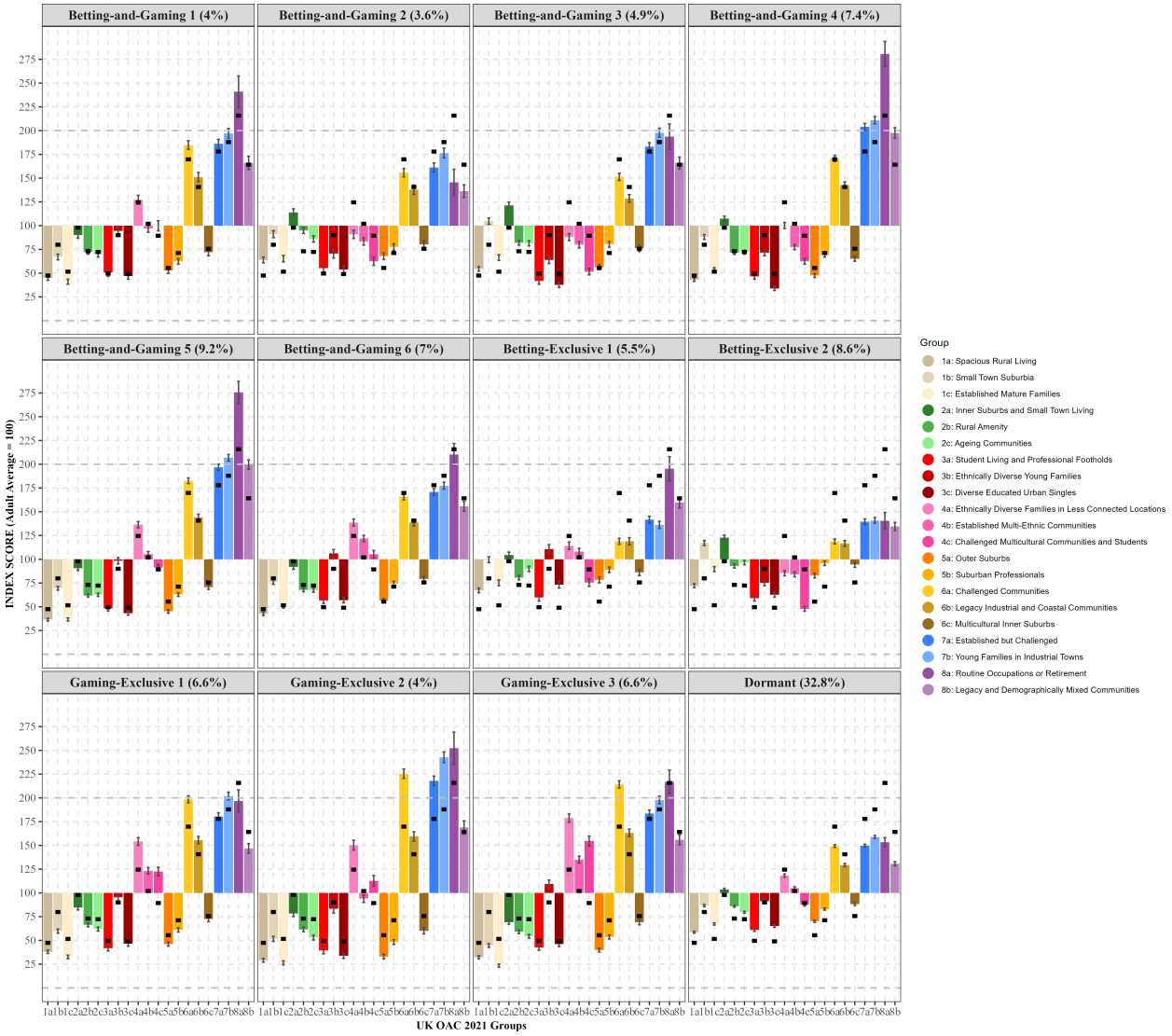

Figure 3: Bar charts illustrating the representation of betting and gaming across 2021 OAC Groups, relative to the adult population of GB [score 100. Black dots denote ISs of all Active Gamblers in the Group. Error bars indicate 95% Confidence Interval (CI)s].

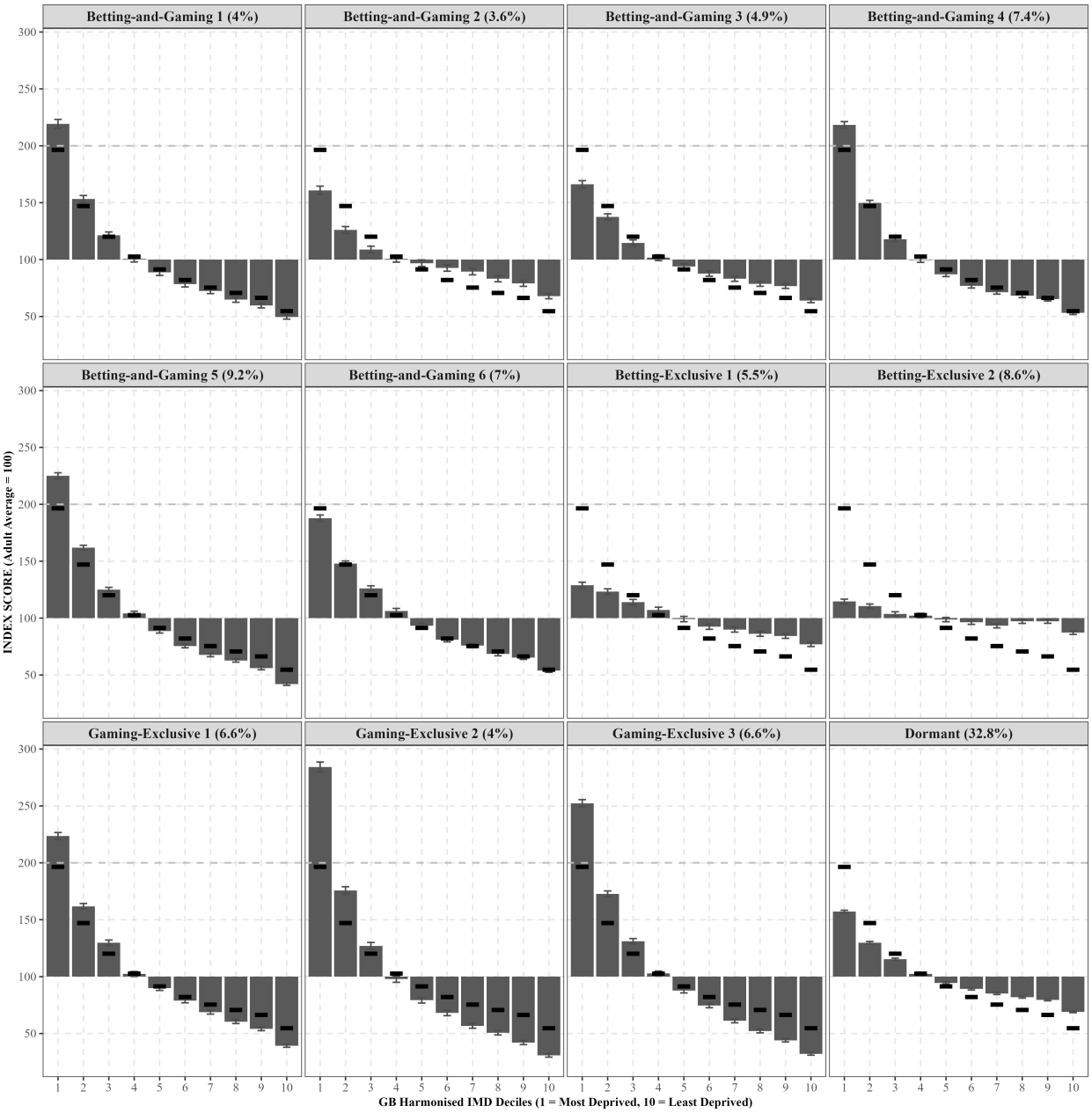

Figure 4: Bar charts illustrating the representation of betting and gaming across 2019 harmonised IMD deciles, relative to the adult population of GB [score 100. Black dots denote ISs of all Active Gamblers in the decile. Error bars indicate 95% CIs].

Table 2

Supergroup	**Group**	**Subgroup**
Active (67.2%)		
	BG: Betting-and-Gaming (36%)	BG1: Engaged New-Age Gamblers (4%)
		BG2: Newcomers with Sporadic and Moderate Patterns of Play (3.6%)
		BG3: Event-Driven Speculative Players (4.9%)
		BG4: Mindful Entertainment Seekers (7.4%)
		BG5: High-Frequency High-Stake Losers (9.2%)
		BG6: Young Occasional Binge Players (7.0%)
	B: Betting-Exclusive (14.1%)	B1: Longstanding Habitual Veterans (5.5%)
		B2: Well-Resourced Hobbyists (8.6%)
	G: Gaming-Exclusive (17.1%)	G1: Episodic High Risk Takers (6.6%)
		G2: Mindful Low-Rollers from Deprived Communities (4.0%)
		G3: High-Stakes Gamers from Deprived Communities (6.6%)
Dormant (32.8%)		
	-	Dormant (32.8%)

Cluster names and distribution statistics of the online gambler classification.

4 Discussions and Conclusions

A data-rich analysis identified distinct gambling profiles, revealing a clear deprivation gradient where some groups are more likely to experience or exacerbate gambling issues. This data rich typology illuminates the geographic and social contexts to gambling, providing a valuable first filter to understand behaviours and the directly measured extent to which they might be conceived as problematic. We see this data rich approach as preferable to surveys of respondent perceived behaviour with much more indirect inference of potentially problematic gambling behaviours. While the findings offer valuable insights, the study relies on data from a single gambling operator and excludes other forms of gambling (e.g., lottery and in-person outlets), limiting generalisability.

In the absence of informed account holder consent, direct targeting of individual account holders using the results of this research would be unethical; instead, our geodemographic approach might be conducive to spatially targeted intervention, for example through profiling of GP registers or practices. Future research should develop active partnerships with providers of all gambling services to provide still more robust and defensible measures of gambling behaviour in order that shared and informed concerns about problem gambling behaviour might be developed alongside responsible policy interventions.

Acknowledgements

This work is funded by the Economic and Social Research Council (ESRC) GeoDS (ref: ES/Z504464/1) in collaboration with GambleAware. The data was supplied by an anonymous provider with no involvement in the research design, and no contractual obligations were in place. This ensures the research remains fully independent of industry influence.

Biographies

Shunya Kimura is a final-year PhD researcher in Social and Geographic Data Science at UCL, funded by the UBEL DTP in collaboration with GambleAware. His research focuses on linking and analysing large-scale behavioural data, with a particular emphasis on the socio-spatial dimensions of online gambling behaviour, to support evidence-based public interventions.

Justin van Dijk is a Lecturer in Social and Geographic Data Science in the Department of Geography at University College London. His research focuses on the analysis and visualisation of large-scale spatial data, with a particular emphasis on socio-spatial inequalities.

Paul A. Longley is Professor of Geographic Information Science at UCL and Director of the UK Geographic Data Service (formally known as Consumer Data Research Centre). His research focuses on the application of geographic information science, with a strong emphasis on the development and deployment of geo-temporal data infrastructures developed from Big or Open Data.

References

Bellemare, M. F., & Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, *82*(1), 50–61. https://doi.org/10.1111/obes.12325

Bigger, B., & Wardle, H. (2024). *Words matter: A language guide for respectful reporting on gambling* (tech. rep.). Glasgow. https://www.grg.scot/resources/2024/08/Words-Matter_main_digital-file.pdf

Braverman, J., & Shaffer, H. J. (2010). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *European Journal of Public Health*, *22*(2), 273–278. https://doi.org/10.1093/eurpub/ckp232

Broda, A., LaPlante, D. A., Nelson, S. E., LaBrie, R. A., Bosworth, L. B., & Shaffer, H. J. (2008). Virtual harm reduction efforts for internet gambling: Effects of deposit limits on actual internet sports gambling behavior. *Harm Reduction Journal*, *5*(1), 27. https://doi.org/10.1186/1477-7517-5-27

Department for Business, Energy and Industrial Strategy. (2021). *Smart data working group: Spring 2021 report* (tech. rep.). assets.publishing.service.gov.uk/media/60c72e058fa8f57ce3773c2d/smart-data-working-group-report-2021.pdf(open%20in%20a%20new%20window).

Fleischmann, M. (2023). Clustergram: Visualization and diagnostics for clusteranalysis. *Journal of Open Source Software*, *8*(89), 5240. https://doi.org/10.21105/joss.05240

Forrest, D., & McHale, I. G. (2022). Patterns of play technical report 2, 109. https://natcen.ac.uk/publications/patterns-play

Gale, C. G., Singleton, A. D., Bates, A. G., & Longley, P. A. (2016). Creating the 2011 area classification for output areas (2011 oac) [Number: 12]. *Journal of Spatial Information Science*, (12), 1–27. https://josis.org/index.php/josis/article/view/66

Ghaharian, K., Abarbanel, B., Phung, D., Puranik, P., Kraus, S., Feldman, A., & Bernhard, B. (2023). Applications of data science for responsible gambling: A scoping review [Publisher: Routledge _eprint: https://doi.org/10.1080/14459795.2022.2135753]. *International Gambling Studies*, *23*(2), 289–312. https://doi.org/10.1080/14459795.2022.2135753

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, gis and neighbourhood targeting* (1st ed.). Wiley. https://web.p.ebscohost.com/ehost/ebookviewer/ebook/bmxlYmtfXzE0MTY4M19fQU41?sid=53b27c6a-4315-4bb2-a67e-41add27be54a@redis&vid=0&format=EB&lpid=lp_i&rid=0

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29. https://doi.org/10.1080/00031305.1992.10475842

Pickering, D., & Blaszczynski, A. (2021). Paid online convenience samples in gambling studies: Questionable data quality. *International Gambling Studies*, *21*(3), 516–536. https://doi.org/10.1080/14459795.2021.1884735

PwC. (2017, August). Remote gambling research interim report on phase ii. https://www.begambleaware.org/sites/default/files/2020-12/gamble-aware_remote-gambling-research_phase-2_pwc-report_august-2017-final.pdf

Rains, T., & Longley, P. (2021). The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services*, *63*, 102650. https://doi.org/10.1016/j.jretconser.2021.102650

Sturgis, P., & Kuha, J. (2021, May). Methodological factors affecting estimates of the prevalence of gambling harm in the united kingdom: A multi-survey study. https://www.begambleaware.org/sites/default/files/2021-05/Methodology_Report_(FINAL_14.05.21).pdf

Wardle, H., Degenhardt, L., Marionneau, V., Reith, G., Livingstone, C., Sparrow, M., Tran, L. T., Biggar, B., Bunn, C., Farrell, M., Kesaite, V., Poznyak, V., Quan, J., Rehm, J., Rintoul, A., Sharma, M., Shiffman, J., Siste, K., Ukhova, D., … Saxena, S. (2024). The lancet public health commission on gambling. *The Lancet Public Health*, *9*(11), e950–e994. https://doi.org/10.1016/S2468-2667(24)00167-1

Watthanacheewakul, L. (2021). Wce 2021. https://www.iaeng.org/publication/WCE2021/WCE2021_pp101-106.pdf

Wyszomierski, J., Longley, P. A., Singleton, A. D., Gale, C., & O'Brien, O. (2024). A neighbourhood output area classification from the 2021 and 2022 uk censuses [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geoj.12550]. *The Geographical Journal*, *190*(2), e12550. https://doi.org/10.1111/geoj.12550