Entropy*

Fabian Gunzinger   Neil Stewart

September 8, 2022

Contents

*We are grateful to Redzo Mujcic and and Zvi Safra for helpful comments. The research was supported by Economic and Social Research Council grant number ES/V004867/1. WBS ethics code: E-414-01-20. Gunzinger: Warwick Business School, fabian.gunzinger@warwick.ac.uk; Stewart: Warwick Business School, neil.stewart@wbs.ac.uk.

1 Introduction

This paper investigates whether they way people spend their money is associated with how much they spend and save.

The question is important because even though around a quarter of adults in the UK and the US are unable to cover irregular expenses like car and medical bills,[1] there is little research that studies why people have low levels of "emergency savings" and test ways to help them build such savings.[2]

Spending profiles are of interest because:

- Research in psychology suggests that disorder is maladaptive and associated with a range of negative outcomes such as impaired executive function (Vernon-Feagans et al. 2016), lower cognitive inhibition (Mittal et al. 2015), and activation of anxiety-related neural circuits (Hirsh et al. 2012).

- In the study of human behaviour, more chaotic behaviour has been found to predict the a higher number of visits to and higher spend in supermarkets (Guidotti et al. 2015), higher calorie intake (Skatova et al. 2019) and financial distress (Muggleton et al. 2020).

Contribution:

- We can think of total savings as the sum of long-term and short-term savings. Long-term savings are savings for retirement, either individually or through an employer-linked pension scheme. These kinds of savings, and especially savings through pension schemes, are well researched. In contrast, there is almost no research on short-term savings, which comprise savings for particular goals such as a new car, a holiday, or a wedding, and emergency savings to have a buffer for unexpected events. (See nest2021supporting for more on emergency savings)

[1] In the UK, 25 percent of adults would be unable to cover an unexpected bill of £300 (Philipps et al. 2021), while in the US, about 30 percent would be unable to cover a $400 bill (Governors of the Federal Reserve System 2022).

[2] In contrast, there is a large body of research on pension savigns. Well-documented behavioural biases that help explain undersaving are, among others, present bias (Laibson 1997, Laibson and Marzilli-Ericson 2019), inertia (Madrian and Shea 2001), over-extrapolation (Choi et al. 2009), and limited self-control and willpower (Thaler and Shefrin 1981, Benhabib and Bisin 2005, Fudenberg and Levine 2006, Loewenstein and O'Donoghue 2004, Gul and Pesendorfer 2001). One danger of viewing low savings mainly as a result of behavioural biases is that while these biases likely do play some role and designing environments and tools to help correct them are thus part of the solution, it is at least conceivable that this is an area where the focus on behaviour-level solutions distracts from an effort to find more effective society-level solutions, a danger inherent in behavioural science research convincingly highlighted in Chater and Loewenstein (2022): if the main problem is that many people are unable to earn enough to save, then the effectiveness of helping them manage their low incomes more effectively pales in comparison with efforts to help them earn more.

- We aim to close this gap. In our data, we cannot distinguish between goal-oriented and emergency savings, so we focus on total short-term savings[3]
- In doing so, we aim to contribute to three literatures:
  - Main 1: Understanding emergency savings behaviour (nest, aspen reports), (Sabat and Gallagher 2019) for sources on short-term savings literature, Colby and Chapman (2013) for lit on savings goals
    * Colby and Chapman (2013) has useful literature review on short-term savings and suggests that subgoals can increase willingness to forego short-amounts in the present because they move the reference point in a prospect-theory framework.
  - Main 2: Understanding effect of behavioural entropy - eliciting useful personality characteristics from large-scale data
  - Also 1: More broadly: part of savings literature (pension literature, savings buffer)
  - Also 2: Use of high-frequency transaction data (itself a sub-literature of use of newly available large-scale datasets)

Structure

- Part I: describtive states of how and when people spend and save
- Part II: define measures that characterise spend profile
- Part III: regression analysis

Literature:

- Savings lit: Lunt and Livingstone (1991) and Oaten and Cheng (2007)

2 Methods

2.1 Dataset description

We use data from Money Dashboard (MDB), a financial management app that allows its users to link accounts from different banks to obtain an integrated view of their finances.[4] The dataset contains more than 500 million transactions made between 2012

[3] MDB allows users to create custom tags and some users use them to indicate the intended use for their savings transactions (e.g. "wedding", "holidays"). But only a very small number of transactions have such tags, and we do not pursue this further.

[4] https://www.moneydashboard.com.

and June 2020 by about 250,000 users, and provides information such as date, amount, and description about the transaction as well as account and user-level information.

The main advantages of the data for the study of consumer financial behaviour are its high frequency, that it is automatically collected and updated and thus less prone to errors and unaffected by biases that bedevil survey measures, and that it offers a view of consumers' entire financial life across all their accounts, rather than just a view of their accounts held at a single bank, provided they added all their accounts to MDB. The main limitation is the non-representativeness of the sample relative to the population as a whole. Financial management apps are known to be used disproportionally by men, younger people, and people of higher socioeconomic status (Carlin et al. 2019). Also, as pointed out in Gelman et al. (2014), a willingness to share financial information with a third party might not only select on demographic characteristics, but also for an increased need for financial management or a higher degree of financial sophistication. Because our analysis does not rely on representativeness, we do not address this.[5]

2.2 Preprocessing and sample selection

We restrict our sample to users for whom we can observe a regular income, can be reasonably sure that they have added all their bank account to MDB, and for whom we observe at least six months of data. Table 1 summarises the sample selection steps we applied to a 1 percent sample of the raw data, associated data losses, and the size of our final sample. A detailed description of the entire data cleaning and selection process is provided in Appendix A.

Table 1: Sample selection

	Users	User-months	Txns	Txns (m£)
Raw sample	25,321	570,959	55,254,410	7,434
Annual income of at least £10k	8,646	181,371	19,140,155	2,588
Income in 2/3 of all observed months	8,646	181,371	19,140,155	2,588
At least one savings account	5,155	117,796	12,945,209	1,866
At least 6 months of data	4,552	116,369	12,820,830	1,847
Monthly debits of at least £200	4,259	109,026	12,176,779	1,758
Five or more monthly current account txns	3,982	102,589	11,397,058	1,635
Minimum level of spend diversity	3,978	102,437	11,384,817	1,633
Complete demographic information	3,333	89,169	9,901,333	1,416
Final sample	3,333	89,169	9,901,333	1,416

[5]For an example of how re-weighing can be used to mitigate the non-representative issue, see Bourquin et al. (2020).

2.3 Dependent variables

Identifying savings transactions: We classify as payments into savings accounts all savings account credits of £5 or more that are not identified as interest payments or automated "save the change" transfers (similarly for debits).[6]

Dummy for savings txn in current month. Motivation: MPS (2018) finds that saving habit is often more important than amount saved.

Gross and net inflows into savings accounts in current month. We winsorise the top end of the distribution at the 1 percent level.

2.4 Spending profile

- Two ways to characterise spend profile: intra-period and inter-period. For proof-of-concept study, we focus on former, and on calendar month as period.[7]

Measures of spending profiles:

- Tag-based entropy
- Smoothed tag-based entropy
- Grocery-shops based entropy
- Composite measure based on PCA from combination of base entropy scores (similar to Eagle et al. (2010)

Tag-based entropy Our variable of interest is spending entropy, a measure of how predictable an individual's spending pattern is at a given point in time, which we interpret more broadly as a measure of the degree to which an individual's life is chaotic. Entropy is a cornerstone of information theory, where it measures the amount of information contained in an event. In the behavioural sciences, behavioural entropy has recently been shown to predict the frequency of grocery visits and the per-capita spend per visit (Guidotti et al. 2015), the amount of calories consumed (Skatova et al. 2019), and the propensity for financial distress (Muggleton et al. 2020).

We calculate spending entropy using the formula proposed by Shannon (1948), which defines entropy as:[8]

$$H = -\sum p_i log(p_i), \tag{1}$$

[6]While standing order transactions are unlikely to be related to entropy in the short-run, we do not exclude such transactions since, best we can tell, the only account for a small fraction of total transactions.

[7]For intra-period characterisation, consistency over time will be particularly interesting. Might be able to capture this using Jensen-Shannon divergence.

[8]Shannon entropy is customarily denoted as H following Shannon's own naming after Ludwig Boltzman's 1872 H-theorem in statistical mechanics, to which it is analagous.

where p_i is the probability that an individual makes a purchase in spending category i, and log is the base 2 logarithm.

We normalise H by $log(N_{SC})$, the entropy of completely random shopping behaviour, so that it takes value between 0 and 1.[9]

The higher the value of entropy, the less predictable an individuals spending pattern.

To calculate entropy scores, we group spending into 9 spending categories (SC) based on the classification used by Lloyds Banking Group as discussed in Muggleton et al. (2020). Also following that paper, we use additive smoothing to calcualte probabilities to avoid taking logs of zeroes in cases where an individual makes no purchases in a given spending category. We thus calculate p_i as:

$$p_i = \frac{\text{Count of purchases in } SC_i + 1}{\text{Count of all purchases} + N_{SC}}, \tag{2}$$

where N_{SC} is the total number of categories.

Issues from imperfect labelling of MDB data:

- Transaction tagging in the MDB data is imperfect: about 20 percent of transactions have no tag.
- This creates two issues for entropy calculation.
- First, entropy scores of high-entropy individuals are biased downwards - relatively more than those of low-entropy individuals. Reason: missing tags are not random: more common transactions such as groceries or take-away purchases are more likely to be tagged ...
- All zero count-cases: ... Solution: require minimum number of txns in two different labels (for all categories we use to calculate entropy). Two to avoid 0 entropy cases that are unlikely to be genuine but probably artefacts of missing labelling.

Auto-tag-based entropy

- Use apriori algorithm
- minsup: minimum number of baskets a pattern is required to appear (else it's dropped)
- In our context: baseket is collection of auto tags with positive txsn counts in a user month, while pattern is pattern of such auto tags.
- Patterns also called 'representative baskets'.

[9] $log(N_{SC})$ is the probability of a completely random shopping pattern because for in this case, for N_{SC} different spending categories, we would have $p_i = 1/N_{SC}$ for each category i so that $H = -N_{SC} p_i log(p_i) = -log(p_i) = log(N_{SC})$.

- Algorithm steps (adapted from guidotti2015behavioural)
  - Identify all patterns (representative baskets)
  - Discard representative baskets that appear in fewer than minsup months we observe for a user.
  - Assign representative basked to each of the user's months.
  - Calculate probabilities of observing a representative basked based on occurrences across all of a user's month. E.g. user with 5 months of data with representative baskets [1, 1, 2, 3, 4] has representative basket probabilitis 2/5 for repr basket 1, and 1/5 for repr baskets 2-4.
  - Calculate user-leven entropy based on probabilities.

Shopping-time based entropy We calculate entropy based on the probability of $(dayofweek, merchant)$ tuples, where we follow Guidotti et al. (2015) and bin *day of week* into *weekends* and *weekday*, to reduce excessive fluctuations. Because banks tend to process weekend transactions on Monday, as shown in Figure 4, we cannot distinguish transactions made on Saturdays or Sundays from those made on Mondays, and thus classify all of them as weekend transactions.

We drop the about 25 percent of transactions for which we cannot identify a merchant. The alternative would be leaving these transactions in the sample and treating "unknown merchant" as a single merchant. But for user-months for which the merchant is unknown for all transactions, this would lead to an entropy score of 0, which is undesireable.

Grocery shop entropy We consider purchases at Tesco, Sainsbury's, Asda, Morrisons, Aldi, Co-op, Lidl, Waitrose, Iceland, and Ocado, which have a combined market share of 96.5 percent.

2.5 Control variables

We classify potential determinants of savings behaviour into *financial behaviours*, *financial planning*, and *individual or household characteristics*, a classification frequently used in policy research on the financial wellbeing (CAN 2019, CFPB 2017, MPS 2018).

Financial behaviour

- Regular savings, dummy for 10 out of last 12 months
- Proportion of purchases paid with credit card. This is only about 6 percent in our final sample, whereas it is 12 percent in the full sample.[10]

[10]Across the UK, the proportion of credit pard purchases is about 17 percent in a typical month (Finance 2021). The proportion in our data is likely lower because the sample is skewed towards more affluent individuals.

- Month total and category spend (category spend for robustness)

Planning

- Regular login, dummy for 1 / month in 10 out of last 12 months. Have login data for about 50 percent of sample, so best to work with full sample once I use it. Implement once I can do that. – not yet implemented –

Individual and household characteristics

- Gender
- Age
- Urban
- Region
- Year income, winsorised at the 1 percent level. We include year-rather than month-income because the latter can be quite variable (e.g. irregular work, changing jobs, etc.), and we assume people to base their consumption on their annual income. (Could test this in appendix: is correlation between annual and spend stronger than between monthly and spend?)
- Regular income, dummy for 10 out of last 12 months
- Month income std – not implemented yet –
- Income current month, dummy for month income > 0
- Has children, imperfect – not yet implemented –
- Index of multiple deprivations from nspl – not implemented yet –
- Received benefits
- Receives pension
- Housing tenure: mortgage, rent, other (owning outright implied)
- Takes out (payday) loan
- Total balance or balance / avg. month spend – not yet implemented –

Table 2: Summary statistics

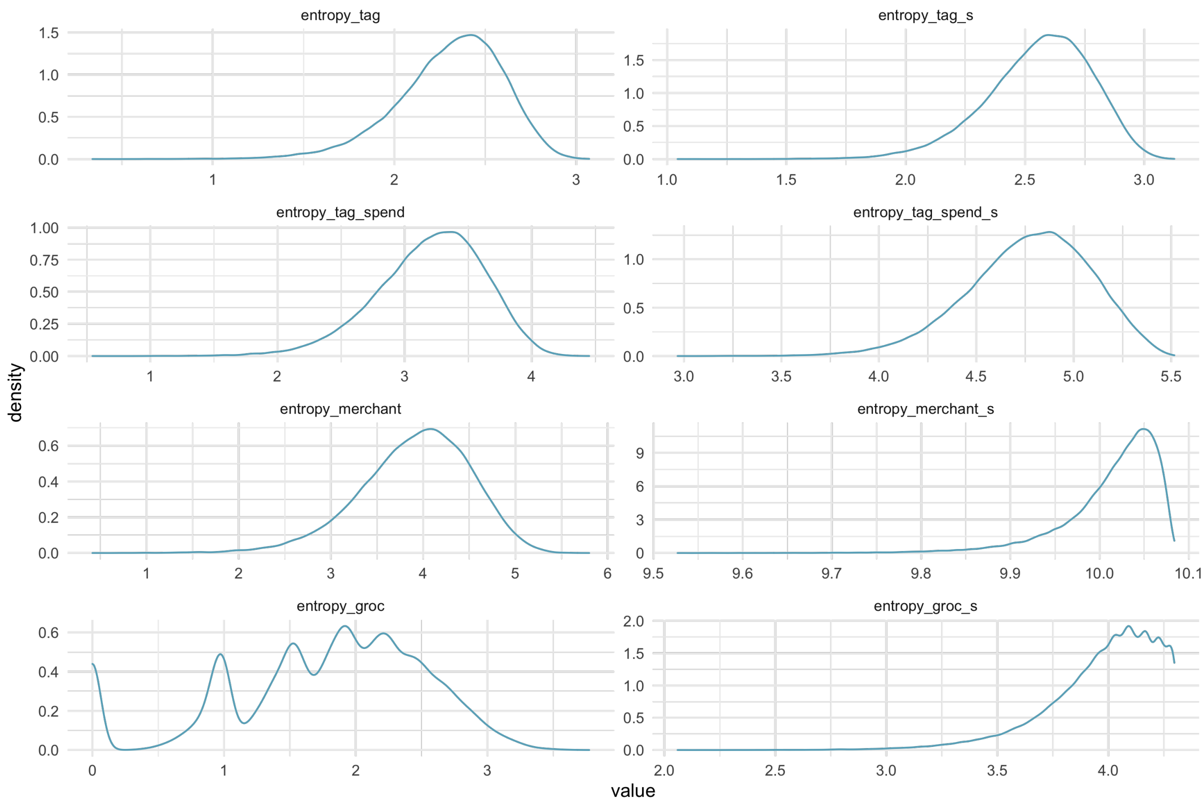
Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
user_id	476,918.200	59,724.570	361,990	426,510	474,750	527,090	589,670
ym	584.110	11.502	555	575	584	593	607
ymn	201,828.900	99.377	201,604	201,712	201,809	201,906	202,008
txns_count	117.829	77.199	1	70	102	145	1,522
txns_volume	24,486.070	95,626.350	202.890	6,008.070	10,672.090	20,875.490	8,973,596.000
month_income	3,210.120	2,642.140	420.343	1,556.612	2,364.224	3,966.723	15,388.180
inflows	852.165	2,875.060	0.000	0.000	0.000	410.000	21,958.480
outflows	850.291	2,781.575	0.000	0.000	0.000	400.000	20,466.230
netflows	−13.629	2,018.053	−10,805.000	0.000	0.000	64.000	10,600.000
netflows_norm	−0.007	0.714	−3.813	0.000	0.000	0.025	3.607
inflows_norm	0.317	0.983	0.000	0.000	0.000	0.181	7.280
outflows_norm	0.328	1.024	0.000	0.000	0.000	0.175	7.595
has_pos_netflows	0.293	0.455	0	0	0	1	1
pos_netflows_norm	0.275	5.083	0.000	0.000	0.000	0.025	712.569
t	0.436	0.496	0	0	0	1	1
tt	−1.981	13.079	−37	−10	−2	6	44
month_spend	3,028.719	3,137.228	200.000	1,267.230	2,102.130	3,538.840	20,442.220
age	37.453	10.817	19	29	35	44	85
is_female	0.398	0.490	0	0	0	1	1
is_urban	0.852	0.355	0	1	1	1	1
has_sa_account	1.000	0.000	1	1	1	1	1
generation_code	2.542	0.684	0	2	3	3	4
new_loan	0.028	0.166	0	0	0	0	1
unemp_benefits	0.000	0.000	0	0	0	0	0
pct_credit	9.959	19.749	0.000	0.000	0.000	9.555	100.000

2.6 Summary statistics

Table 2 provides summary statistics.

Figure 1

Figure 1: Entropy distributions



Notes:

2.7 Model specification

We estimate models of the form:

$$s_{i,t} = \alpha_i + \lambda_t + \beta H_{i,t} + x'_{i,t}\delta + \epsilon_{i,t}, \tag{3}$$

where $s_{i,t}$ is an indicator variable equal to one if individual i made one or more transfers to any of their savings account in month t and zero otherwise, H_{it} is i's spending entropy in month t, $x_{i,t}$ a vector of control variables, α_i an individual fixed effect, λ_t a calendar month fixed effect, and $\epsilon_{i,t}$ the error term.

Issues to think about:

- Unbalanced panel is not random - people using MDB for longer are different. Should we just use first x months for every user? E.g. first year?

3 Results

3.1 Main results

3.2 Tables

Table 3: has_sa_inflows on unsmoothed entropy

Dependent Variable:	Has savings							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (9 cats)	0.037***				0.014***			
	[0.030; 0.043]				[0.004; 0.023]			
Entropy (48 cats)		0.046***				0.031***		
		[0.039; 0.052]				[0.020; 0.042]		
Entropy (merchant)			0.058***				0.024***	
			[0.052; 0.064]				[0.014; 0.035]	
Entropy (groceries)				0.027***				0.009***
				[0.023; 0.030]				[0.005; 0.014]
Paid with credit (%)	-0.002***	-0.002***	-0.002***	-0.002***	0.000	0.000	0.000	0.000
	[-0.002; -0.002]	[-0.002; -0.002]	[-0.002; -0.002]	[-0.002; -0.002]	[-0.000; 0.000]	[-0.000; 0.000]	[-0.000; 0.000]	[-0.000; 0.000]
Month spend	-0.000	-0.002*	-0.004***	-0.001	0.010***	0.010***	0.009***	0.010***
	[-0.002; 0.002]	[-0.004; 0.000]	[-0.006; -0.002]	[-0.003; 0.001]	[0.008; 0.013]	[0.007; 0.012]	[0.007; 0.012]	[0.008; 0.013]
Urban	0.009*	0.010**	0.008*	0.008*	-0.026	-0.024	-0.023	-0.029
	[-0.001; 0.018]	[0.001; 0.019]	[-0.001; 0.017]	[-0.001; 0.017]	[-0.233; 0.181]	[-0.235; 0.187]	[-0.234; 0.187]	[-0.235; 0.177]
Month income	-0.020***	-0.021***	-0.020***	-0.019***	0.002	0.002	0.002	0.002
	[-0.023; -0.018]	[-0.023; -0.018]	[-0.023; -0.018]	[-0.021; -0.016]	[-0.005; 0.009]	[-0.005; 0.009]	[-0.005; 0.009]	[-0.004; 0.009]
Has income in month	0.126***	0.124***	0.119***	0.120***	0.045***	0.044***	0.043***	0.045***
	[0.097; 0.155]	[0.094; 0.153]	[0.090; 0.148]	[0.091; 0.149]	[0.015; 0.075]	[0.014; 0.074]	[0.013; 0.073]	[0.016; 0.075]
Income variability	-0.024***	-0.024***	-0.023***	-0.023***	-0.003	-0.003	-0.003	-0.003
	[-0.028; -0.019]	[-0.029; -0.019]	[-0.028; -0.019]	[-0.027; -0.018]	[-0.012; 0.006]	[-0.012; 0.006]	[-0.012; 0.006]	[-0.012; 0.006]
Rent payment	-0.011**	-0.015***	-0.011**	-0.015***	0.028***	0.024**	0.027***	0.027***
	[-0.021; -0.001]	[-0.025; -0.005]	[-0.021; -0.001]	[-0.025; -0.005]	[0.008; 0.047]	[0.005; 0.044]	[0.008; 0.046]	[0.008; 0.047]
Mortgage payment	-0.021***	-0.029***	-0.025***	-0.020***	0.022*	0.017	0.020	0.022*
	[-0.030; -0.012]	[-0.037; -0.020]	[-0.034; -0.017]	[-0.029; -0.011]	[-0.002; 0.046]	[-0.007; 0.041]	[-0.004; 0.044]	[-0.002; 0.046]
Loan repayment	0.017***	0.013***	0.011***	0.020***	0.008	0.005	0.006	0.008
	[0.010; 0.024]	[0.006; 0.020]	[0.004; 0.018]	[0.013; 0.027]	[-0.008; 0.023]	[-0.011; 0.020]	[-0.009; 0.022]	[-0.007; 0.024]
Benefits	0.022***	0.018***	0.018**	0.016**	-0.005	-0.007	-0.007	-0.007
	[0.008; 0.035]	[0.004; 0.031]	[0.004; 0.031]	[0.003; 0.030]	[-0.043; 0.033]	[-0.045; 0.031]	[-0.045; 0.031]	[-0.045; 0.031]
(Intercept)	0.316***	0.321***	0.325***	0.330***				
	[0.285; 0.346]	[0.291; 0.352]	[0.294; 0.355]	[0.300; 0.361]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	83,935	83,935	83,935	83,935	83,935	83,935	83,935	83,935
R^2	0.01875	0.01932	0.02135	0.01996	0.42806	0.42841	0.42828	0.42813
Within R^2					0.00280	0.00341	0.00317	0.00292

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Spend and income variables are in £'000.

Table 4: has_sa_inflows on smoothed entropy

Dependent Variable:	Has savings							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (9 cats, smooth)	-0.001 [-0.005; 0.003]				-0.004 [-0.010; 0.002]			
Entropy (48 cats, smooth)		-0.028*** [-0.032; -0.023]				-0.016*** [-0.023; -0.008]		
Entropy (merchant, smooth)			-0.027*** [-0.031; -0.023]				-0.017*** [-0.025; -0.010]	
Entropy (groceries, smooth)				-0.017*** [-0.021; -0.014]				-0.010*** [-0.016; -0.004]
Paid with credit (%)	-0.002*** [-0.002; -0.002]	-0.002*** [-0.002; -0.002]	-0.002*** [-0.002; -0.002]	-0.002*** [-0.002; -0.002]	0.000 [-0.000; 0.000]	-0.000 [-0.000; 0.000]	-0.000 [-0.000; 0.000]	0.000 [-0.000; 0.000]
Month spend	0.001 [-0.001; 0.003]	-0.002* [-0.004; 0.000]	-0.003*** [-0.005; -0.001]	-0.000 [-0.002; 0.002]	0.011*** [0.008; 0.013]	0.009*** [0.007; 0.012]	0.009*** [0.006; 0.011]	0.010*** [0.007; 0.012]
Urban	0.008* [-0.001; 0.017]	0.004 [-0.006; 0.013]	0.004 [-0.005; 0.013]	0.007 [-0.003; 0.016]	-0.028 [-0.233; 0.177]	-0.027 [-0.231; 0.178]	-0.027 [-0.231; 0.178]	-0.030 [-0.235; 0.175]
Month income	-0.019*** [-0.022; -0.017]	-0.020*** [-0.023; -0.018]	-0.021*** [-0.024; -0.018]	-0.019*** [-0.022; -0.017]	0.002 [-0.004; 0.009]	0.002 [-0.005; 0.009]	0.001 [-0.006; 0.008]	0.002 [-0.005; 0.009]
Has income in month	0.127*** [0.098; 0.156]	0.123*** [0.093; 0.152]	0.125*** [0.095; 0.154]	0.124*** [0.094; 0.153]	0.046*** [0.016; 0.076]	0.043*** [0.013; 0.073]	0.044*** [0.015; 0.074]	0.045*** [0.015; 0.074]
Income variability	-0.024*** [-0.029; -0.020]	-0.024*** [-0.029; -0.020]	-0.025*** [-0.029; -0.020]	-0.024*** [-0.029; -0.019]	-0.003 [-0.012; 0.006]	-0.003 [-0.012; 0.006]	-0.003 [-0.012; 0.005]	-0.003 [-0.012; 0.005]
Rent payment	-0.014*** [-0.024; -0.004]	-0.014*** [-0.024; -0.004]	-0.014*** [-0.024; -0.004]	-0.014*** [-0.024; -0.004]	0.027*** [0.008; 0.047]	0.028*** [0.009; 0.048]	0.028*** [0.009; 0.048]	0.028*** [0.008; 0.047]
Mortgage payment	-0.019*** [-0.028; -0.011]	-0.013*** [-0.022; -0.005]	-0.015*** [-0.024; -0.006]	-0.018*** [-0.026; -0.009]	0.022* [-0.002; 0.046]	0.023* [-0.001; 0.047]	0.023* [-0.001; 0.046]	0.022* [-0.002; 0.046]
Loan repayment	0.022*** [0.015; 0.029]	0.024*** [0.017; 0.031]	0.024*** [0.017; 0.031]	0.023*** [0.016; 0.030]	0.009 [-0.007; 0.025]	0.009 [-0.006; 0.025]	0.009 [-0.007; 0.024]	0.009 [-0.007; 0.024]
Benefits	0.017** [0.004; 0.031]	0.015** [0.001; 0.028]	0.013* [-0.001; 0.026]	0.017** [0.003; 0.030]	-0.006 [-0.044; 0.032]	-0.007 [-0.045; 0.031]	-0.008 [-0.046; 0.030]	-0.006 [-0.044; 0.032]
(Intercept)	0.323*** [0.293; 0.354]	0.331*** [0.301; 0.362]	0.339*** [0.308; 0.369]	0.330*** [0.299; 0.360]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	83,935	83,935	83,935	83,935	83,935	83,935	83,935	83,935
R^2	0.01726	0.01924	0.01924	0.01844	0.42798	0.42823	0.42832	0.42815
Within R^2					0.00265	0.00309	0.00325	0.00294

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Spend and income variables are in £'000.

3.3 Exploration

Table 5: Entropy exploration

Dependent Variable:		Has savings	
Model:	(1)	(2)	(3)
Variables			
Entropy (48 cats)		0.029***	
		[0.013; 0.044]	
Entropy (48 cats, smooth)			0.004
			[-0.008; 0.016]
Spend txns	0.001***	0.001***	0.001***
	[0.000; 0.001]	[0.001; 0.001]	[0.000; 0.001]
Unique categories	0.004***	0.001	0.004***
	[0.002; 0.006]	[-0.002; 0.003]	[0.002; 0.006]
Paid with credit (%)	-0.000	-0.000	-0.000
	[-0.001; 0.000]	[-0.001; 0.000]	[-0.001; 0.000]
Month spend	0.006***	0.005***	0.006***
	[0.003; 0.008]	[0.003; 0.008]	[0.003; 0.008]
Urban	-0.022	-0.019	-0.021
	[-0.230; 0.187]	[-0.230; 0.192]	[-0.230; 0.187]
Month income	0.000	-0.000	0.000
	[-0.007; 0.007]	[-0.007; 0.006]	[-0.007; 0.007]
Has income in month	0.038**	0.038**	0.038**
	[0.008; 0.068]	[0.008; 0.068]	[0.009; 0.068]
Income variability	-0.003	-0.003	-0.003
	[-0.012; 0.006]	[-0.012; 0.005]	[-0.012; 0.005]
Rent payment	0.023**	0.024**	0.023**
	[0.004; 0.043]	[0.004; 0.043]	[0.004; 0.043]
Mortgage payment	0.015	0.015	0.015
	[-0.009; 0.039]	[-0.009; 0.039]	[-0.009; 0.039]
Loan repayment	0.003	0.003	0.003
	[-0.013; 0.019]	[-0.012; 0.019]	[-0.013; 0.019]
Benefits	-0.010	-0.011	-0.010
	[-0.048; 0.028]	[-0.049; 0.027]	[-0.048; 0.028]
Fixed-effects			
User id	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes
Fit statistics			
Observations	83,935	83,935	83,935
R^2	0.42912	0.42929	0.42913
Within R^2	0.00464	0.00494	0.00465

Clustered (User id) co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Spend and income variables are in £'000.

Table 6: Components exploration

Dependent Variable:	Has savings			
Model:	(1)	(2)	(3)	(4)
Variables				
Entropy (48 cats)		0.032***		0.044***
		[0.021; 0.042]		[0.024; 0.064]
Average spend			0.000*	0.000**
			[-0.000; 0.000]	[0.000; 0.000]
Cnz			0.006***	0.001
			[0.004; 0.007]	[-0.002; 0.003]
Counts std all			0.004**	0.011***
			[0.001; 0.008]	[0.006; 0.016]
Paid with credit (%)	-0.000	-0.000	-0.000**	-0.000**
	[-0.001; 0.000]	[-0.001; 0.000]	[-0.001; -0.000]	[-0.001; -0.000]
Month spend	0.020***	0.019***	0.015***	0.014***
	[0.018; 0.023]	[0.017; 0.022]	[0.011; 0.019]	[0.010; 0.018]
Urban	-0.039	-0.037	-0.037	-0.036
	[-0.230; 0.152]	[-0.234; 0.159]	[-0.233; 0.158]	[-0.234; 0.163]
Month income	0.004	0.003	0.003	0.002
	[-0.003; 0.011]	[-0.004; 0.010]	[-0.004; 0.009]	[-0.005; 0.009]
Has income in month	0.039**	0.037**	0.032**	0.031*
	[0.007; 0.071]	[0.005; 0.069]	[0.000; 0.064]	[-0.000; 0.063]
Income variability	0.003	0.003	0.003	0.003
	[-0.006; 0.012]	[-0.006; 0.012]	[-0.006; 0.011]	[-0.006; 0.011]
Rent payment	0.028***	0.025***	0.023**	0.023**
	[0.009; 0.047]	[0.006; 0.043]	[0.004; 0.041]	[0.004; 0.042]
Mortgage payment	0.008	0.003	-0.000	-0.000
	[-0.016; 0.033]	[-0.022; 0.027]	[-0.025; 0.024]	[-0.025; 0.024]
Loan repayment	0.004	-0.000	-0.003	-0.002
	[-0.012; 0.020]	[-0.016; 0.016]	[-0.019; 0.013]	[-0.018; 0.014]
Benefits	-0.004	-0.005	-0.007	-0.008
	[-0.041; 0.033]	[-0.043; 0.032]	[-0.044; 0.030]	[-0.045; 0.030]
Fixed-effects				
User id	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes
Fit statistics				
Observations	91,644	91,644	91,644	91,644
R^2	0.41730	0.41776	0.41807	0.41833
Within R^2	0.00782	0.00860	0.00914	0.00957

Clustered (User id) co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 7: Entropy exploration

Dependent Variable:	Entropy (48 cats)				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Entropy (48 cats, smooth)	0.274***				0.539***
	[0.264; 0.285]				[0.528; 0.550]
Spend txns		0.001***			0.010***
		[0.001; 0.002]			[0.010; 0.010]
Unique categories			0.097***		0.081***
			[0.096; 0.099]		[0.079; 0.082]
Category count variation				-0.000***	-0.000***
				[-0.000; -0.000]	[-0.000; -0.000]
Fixed-effects					
User id	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	83,935	83,935	83,935	83,935	83,935
R^2	0.66423	0.59311	0.77939	0.61808	0.92879
Within R^2	0.17916	0.00530	0.46067	0.06634	0.82592

Clustered (User id) co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Spend and income variables are in £'000.

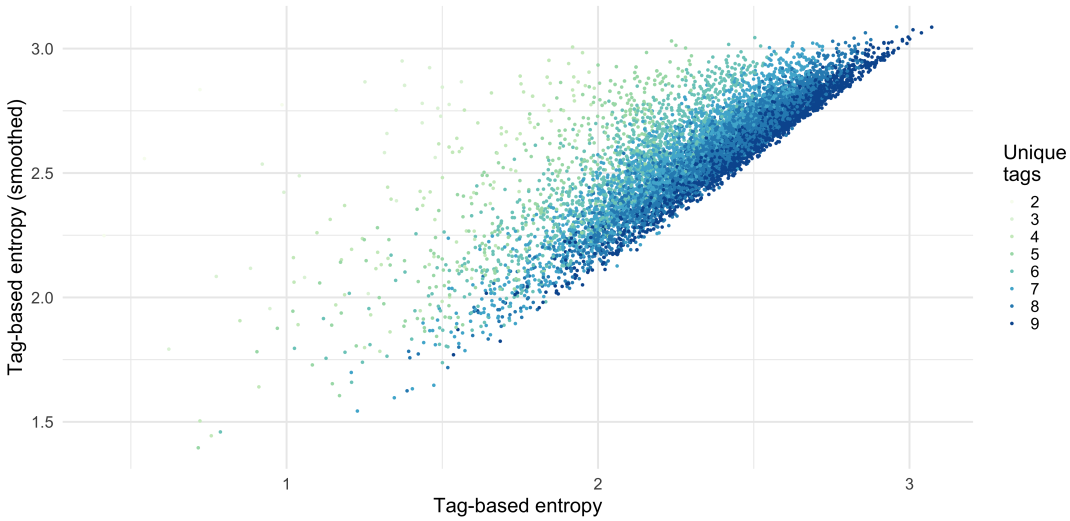
Table 8: Entropy exploration

Dependent Variable:	Entropy (48 cats)				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Entropy (48 cats, smooth)	0.212***				0.530***
	[0.208; 0.216]				[0.526; 0.533]
Spend txns		0.003***			0.010***
		[0.003; 0.003]			[0.010; 0.010]
Unique categories			0.098***		0.080***
			[0.097; 0.098]		[0.080; 0.081]
Category count variation				-0.000***	-0.000***
				[-0.000; -0.000]	[-0.000; -0.000]
(Intercept)	0.501***	0.213***	-1.079***	0.508***	-1.228***
	[0.497; 0.504]	[0.205; 0.221]	[-1.088; -1.069]	[0.503; 0.512]	[-1.232; -1.223]
Fit statistics					
Observations	83,935	83,935	83,935	83,935	83,935
R^2	0.10447	0.04296	0.56104	0.02614	0.88513
Adjusted R^2	0.10446	0.04295	0.56104	0.02613	0.88513

IID co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Spend and income variables are in £'000.

Figure 2: Entropy and unique tags



4 Discussion

References

Benhabib, Jess and Alberto Bisin (2005). "Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions". In: *Games and economic Behavior* 52.2, pp. 460–492.

Bourquin, Pascale, Isaac Delestre, Robert Joyce, Imran Rasul, and Tom Walters (2020). "The effects of coronavirus on household finances and financial distress". In: *IFS Briefing Note BN298*.

CAN, Commonwealth Bank of Australia (2019). "Improving the Financial Wellbeing of Australians". Tech. rep. URL: https://www.commbank.com.au/content/dam/caas/newsroom/docs/CWM0375_Financial%20Wellbeing%20Report_v4.pdf.

Carlin, Bruce, Arna Olafsson, and Michaela Pagel (2019). "Generational Differences in Managing Personal Finances". In: *AEA Papers and Proceedings*. Vol. 109, pp. 54–59.

CFPB, Consumer Financial Protection Bureau (2017). "Financial Well-being in America". Tech. rep. URL: https://www.consumerfinance.gov/data-research/research-reports/financial-well-being-america/.

Chater, Nick and George Loewenstein (2022). "The i-frame and the s-frame: How focusing on the individual-level solutions has led behavioral public policy astray". In: *Available at SSRN 4046264*.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick (2009). "Reinforcement learning and savings behavior". In: *The Journal of finance* 64.6, pp. 2515–2534.

Colby, Helen and Gretchen B Chapman (2013). "Savings, subgoals, and reference points". In:

Eagle, Nathan, Michael Macy, and Rob Claxton (2010). "Network diversity and economic development". In: *Science* 328.5981, pp. 1029–1031.

Finance, UK (2021). "Card Spending Update for November 2021". Tech. rep. URL: https://www.ukfinance.org.uk/sites/default/files/uploads/Data%20(XLS%20and%20PDF)/Card-Spending-Update-November-2021-final.pdf.

Fudenberg, Drew and David K Levine (2006). "A dual-self model of impulse control". In: *American economic review* 96.5, pp. 1449–1476.

Gelman, Michael, Shachar Kariv, Matthew D Shapiro, Dan Silverman, and Steven Tadelis (2014). "Harnessing naturally occurring data to measure the response of spending to income". In: *Science* 345.6193, pp. 212–215.

Governors of the Federal Reserve System, Board of (2022). "Economic Well-Being of U.S. Households in 2021". Tech. rep.

Guidotti, Riccardo, Michele Coscia, Dino Pedreschi, and Diego Pennacchioli (2015). "Behavioral entropy and profitability in retail". In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 1–10.

Gul, Faruk and Wolfgang Pesendorfer (2001). "Temptation and self-control". In: *Econometrica* 69.6, pp. 1403–1435.

Hirsh, Jacob B, Raymond A Mar, and Jordan B Peterson (2012). "Psychological entropy: a framework for understanding uncertainty-related anxiety." In: *Psychological review* 119.2, p. 304.

Krumme, Coco, Alejandro Llorente, Manuel Cebrian, Alex Pentland, and Esteban Moro (2013). "The predictability of consumer visitation patterns". In: *Scientific reports* 3.1, pp. 1–5.

Laibson, David (1997). "Golden eggs and hyperbolic discounting". In: *The Quarterly Journal of Economics* 112.2, pp. 443–478.

Laibson, David and Keith Marzilli-Ericson (2019). "Intertemporal choice". In: *Handbook of Behavioral Economics* 2.

Loewenstein, George and Ted O'Donoghue (2004). "Animal spirits: Affective and deliberative processes in economic behavior". In: *Available at SSRN 539843*.

Lunt, Peter K and Sonia M Livingstone (1991). "Psychological, social and economic determinants of saving: Comparing recurrent and total savings". In: *Journal of economic Psychology* 12.4, pp. 621–641.

Madrian, Brigitte C and Dennis F Shea (2001). "The power of suggestion: Inertia in 401 (k) participation and savings behavior". In: *The Quarterly journal of economics* 116.4, pp. 1149–1187.

Mittal, Chiraag, Vladas Griskevicius, Jeffry A Simpson, Sooyeon Sung, and Ethan S Young (2015). "Cognitive adaptations to stressful environments: When childhood adversity enhances adult executive function." In: *Journal of personality and social psychology* 109.4, p. 604.

MPS, Money and Pension Service (2018). "Building the Financial Capability of UK Adults". Tech. rep. URL: https://moneyandpensionsservice.org.uk/2019/02/06/adult-financial-capability-building-the-financial-capability-of-uk-adults-survey/.

Muggleton, Naomi K, Edika G Quispe-Torreblanca, David Leake, John Gathergood, and Neil Stewart (2020). "Evidence from mass-transactional data that chaotic spending behaviour precedes consumer financial distress". Tech. rep. DOI: 10.31234/osf.io/qabgm. URL: psyarxiv.com/qabgm.

Oaten, Megan and Ken Cheng (2007). "Improvements in self-control from financial monitoring". In: *Journal of Economic Psychology* 28.4, pp. 487–501.

Philipps, Jo, Annick Kuipers, and Will Sandbrook (2021). "Supporting emergency savings: early learnings of the employee experience of workplace sidecar savings". Tech. rep.

Sabat, Jorge and Emily Gallagher (2019). "Rules of thumb in household savings decisions: Estimation using threshold regression". In: *Available at SSRN 3455696*.

Shannon, Claude Elwood (1948). “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3, pp. 379–423.

Skatova, Anya, Neil Stewart, Edward Flavahan, and James Goulding (2019). “Those Whose Calorie Consumption Varies Most Eat Most”. In:

Thaler, Richard H and Hersh M Shefrin (1981). “An economic theory of self-control”. In: *Journal of political Economy* 89.2, pp. 392–406.

Vernon-Feagans, Lynne, Michael Willoughby, and Patricia Garrett-Peters (2016). “Predictors of behavioral regulation in kindergarten: Household chaos, parenting, and early executive functions.” In: *Developmental psychology* 52.3, p. 430.

A Data

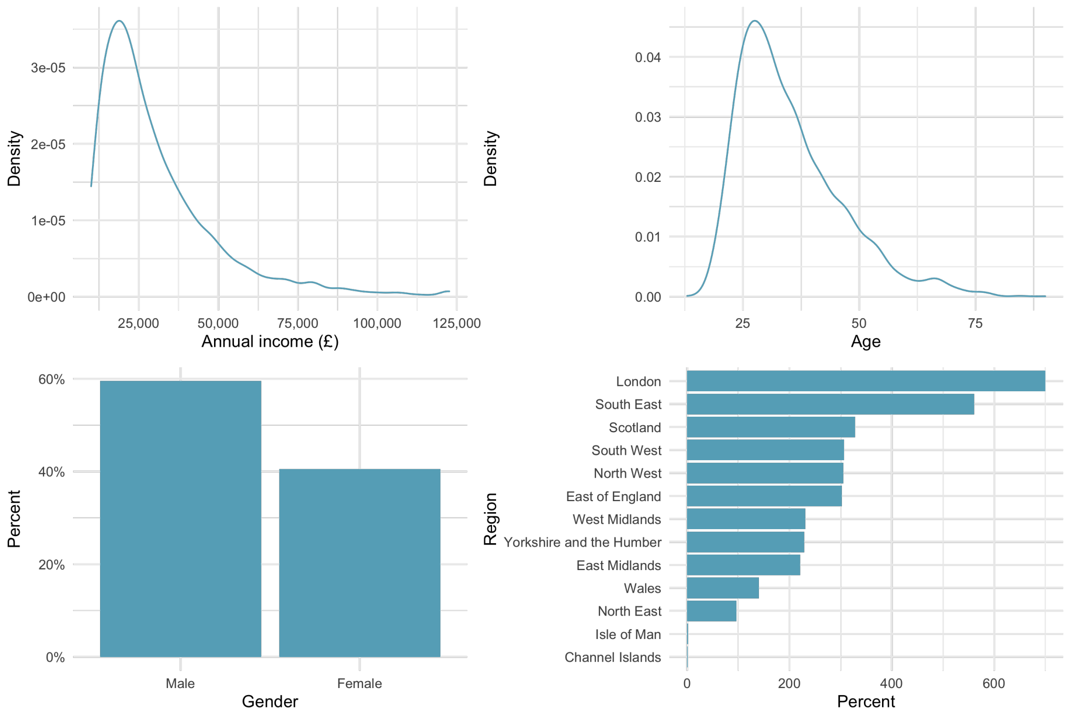
A.1 Data preprocessing

Preprocessing steps (provide detailes and links to relevant code files)

- Duplicates handling.
- We trim all variables at the 1-percent level on the upper end of the distribution for variables that take non-negative values only and on both ends of the distribution for all other variables. We trim (replace outliers with missing values) rather than winsorise (replace outliers with the cutoff percentile value) because we believe that outliers result from errors in the data rather than represent genuine information.
- Actually, we don’t do either of the above. With the harsher selection methods, the statistics are very reasonable, which, if anything, would suggest using winsorizing. However, [this](https://blogs.sas.com/content/iml/2017/02/08/winsorization-good-bad-and-ugly.html) article convincingly argues that we shouldn’t do that in our case.

A.2 Variabel description

Figure 3: Demographic characteristics of Money Dashboard users



Exploration of control variables here (e.g. like jpmorgan2019weathering for income stability)

A.3 Data issues

Bourquin et al. (2020) argue that because some of the accounts in the data will be joint accounts, units of observations should be tought of as "households" rather than "users". We do not agree that this is the most prudent approach. The validity of thinking of units as households depends on the proportion of users in the data who add joint accounts and on the proportion of transactions – out of a user's total number of transactions – additionally observed as a result. Given that the sample is skewed towards younger individuals we think it is unlikely that a majority of them has added joint accounts. Furthermore, it seems reasonable to assume that in most cases, joint accounts are mainly used for common household expenditures similar that are similar to those of a single user (albeit in higher amounts), and are thus unlikely to alter the observed spending profile much. Thus, we think of units of observations as individuals, not households.

Some accounts might be business accounts. Using versions of the algorightms used by Bourquin et al. (2020) to identify such accounts showed, however, that such accounts only make up a tiny percentage of overall accounts and would not influence our results. We thus do not exclude them.

B Entropy

In equation 1 we have defined entropy as $H = -\sum p_i log(p_i)$ and pointed out that it can loosely be interpreted as the predictability of an individual's spending behaviour. In this section, we provide a more detailed discussion of the formula.

The building blocks of entropy is the information content of a single event. The key intuition Shannon (1948) aimed to capture was that learning of the occurrence of a low-probability event is more informative than learning of the occurrence of a high-probability event. The information of an event $I(E)$ is thus inversely proportional to is probability $p(E)$. One way to capture this would be to define the information of event E as $I(E) = \frac{1}{p(E)}$. Yet this implied that an event that is certain to occur had information 1, when it would make sense to have information 0. To remedy this (and also satisfy additional desireable characteristics of an information function), we can can use the log of the expression. Hence, the information of event E, often called *Shannon information*, *self-information*, or just *information*, is defined as:

$$I(E) = log\left(\frac{1}{p(E)}\right) = -log(p(E)) \tag{4}$$

Entropy, often called *Information entropy*, *Shannon entropy*, or just *entropy*, is the information of a random variable and captures the expected amount of information of an event drawn at random from the probability distribution of the random variable. It is calcualted as:

$$H(X) = -\sum_x p(x) \times log(p(x)) = \sum_x p(x)I(x) = \mathbb{E}I(x). \tag{5}$$

todo: Discuss link to spending behaviour

B.1 Entropy calculation

Entropy can be calculated along a number of dimensions.

- Category-based vs time-based vs category-time based (Guidotti et al. 2015, Krumme et al. 2013)
- Count-based vs value-based
- Intratemporal vs intertemporal (Krumme et al. 2013)

    - Based on behaviour within a given time period or changes in behaviour across time periods.

Desireable features of entropy variable:

- Based on a large enough number of categories so that spend on many of them can reasonably be interpreted as chaotic (the 9 LBG tags seem insufficient for this, especially because most of them are vital life expenses). Use of auto tags or merchant seems preferable.

-

C Effect of entropy on overdraft fees

- Table 9 replicates results from Muggleton et al. (2020) Tables S20 (Columns 1 and 2) and Table S40 (columns 3 and 4).
- Similar to their results, higher entropy is positively related to negative financial outcomes.
- Yet, contrary to their findings, the effect becomes stronger in the presence of fixed effects.
- Also, in contrast to above findings, using non-smoothed entropy doesn't reverse the effect but strengthen it.
- Differences to Muggleton et al. (2020) could be due to a number of factors: we use a different dataset, a different and self-selected sample of people that is skewed towards younger and wealthier individuals, as well as a longer panel.

D Spending patterns

Figure 4

Table 9: Effect of entropy on overdraft fees

Dependent Variable:	has_od_fees			
Model:	(1)	(2)	(3)	(4)
Variables				
entropy_tag_sz	0.011***		0.010***	
	[0.007; 0.014]		[0.006; 0.015]	
entropy_tag_z		0.081***		0.040***
		[0.075; 0.086]		[0.032; 0.048]
Spend communication	0.706***	0.596***	0.122***	0.091**
	[0.657; 0.755]	[0.546; 0.645]	[0.051; 0.194]	[0.020; 0.162]
Spend finance	0.035***	0.028***	0.015***	0.013***
	[0.029; 0.041]	[0.022; 0.033]	[0.007; 0.023]	[0.005; 0.021]
Spend hobbies	0.026	-0.078***	0.052*	0.019
	[-0.027; 0.079]	[-0.131; -0.025]	[-0.003; 0.108]	[-0.037; 0.076]
Spend household	-0.002	0.0003	0.003	0.003*
	[-0.005; 0.001]	[-0.003; 0.003]	[-0.0010; 0.007]	[-0.0005; 0.007]
Spend other	0.024***	0.010**	-0.003	-0.008**
	[0.015; 0.032]	[0.002; 0.018]	[-0.011; 0.004]	[-0.015; -0.0003]
Spend motor	0.103***	0.014	0.019	-0.010
	[0.073; 0.132]	[-0.016; 0.044]	[-0.021; 0.059]	[-0.050; 0.030]
Spend retail	0.006	-0.018**	-0.006	-0.015**
	[-0.008; 0.020]	[-0.033; -0.004]	[-0.018; 0.006]	[-0.027; -0.003]
Spend services	-0.010***	0.002	-0.005	-0.003
	[-0.017; -0.003]	[-0.005; 0.009]	[-0.012; 0.002]	[-0.010; 0.003]
Spend travel	-0.033***	-0.045***	-0.0005	-0.005
	[-0.042; -0.023]	[-0.054; -0.035]	[-0.007; 0.006]	[-0.011; 0.002]
Female	0.041***	0.036***	-0.342	-0.317
	[0.035; 0.047]	[0.030; 0.042]	[-78,543.0; 78,542.3]	[-78,455.4; 78,454.7]
Age	-0.002***	-0.002***		
	[-0.003; -0.002]	[-0.003; -0.002]		
Year income	-0.002***	-0.002***	0.0002	0.0001
	[-0.002; -0.001]	[-0.002; -0.001]	[-0.0003; 0.0006]	[-0.0003; 0.0006]
(Intercept)	0.274***	0.266***		
	[0.263; 0.284]	[0.255; 0.276]		
Fixed-effects				
User id			Yes	Yes
Calendar month			Yes	Yes
Fit statistics				
Observations	89,169	89,169	89,169	89,169
R^2	0.02481	0.03254	0.63854	0.63944
Within R^2			0.00182	0.00431

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 4: Spending behaviour

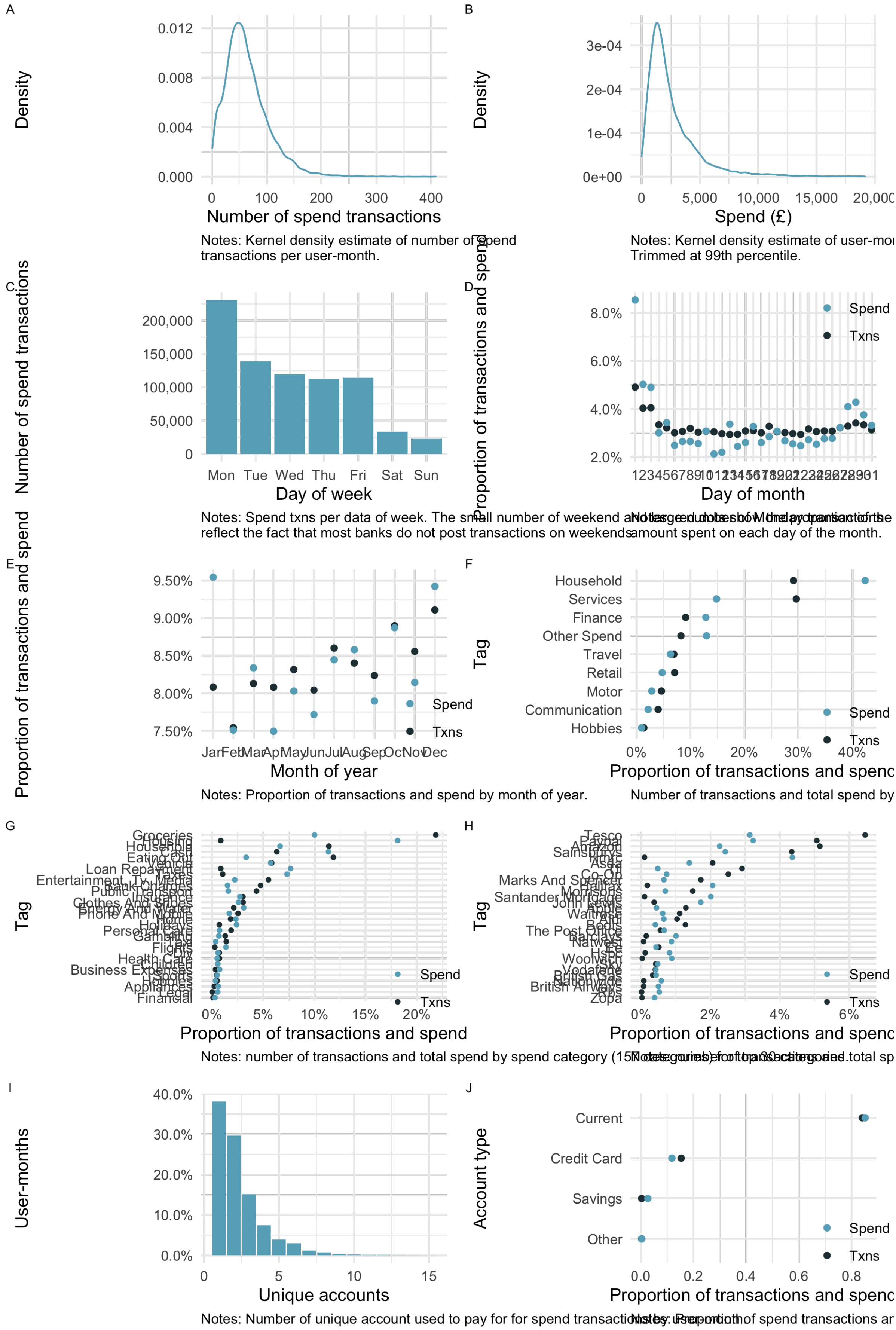


Figure 5: Savings behaviour