Entropy*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

March 31, 2022

Abstract

...

Contents

*This research was supported by Economic and Social Research Council grant number ES/V004867/1. WBS ethics code: E-414-01-20.

1 Introduction

Overall plan for paper

- Question: is the way people spend related to how they save? (I.e. is there a relationship between spend profiles and savings beheaviour?)
- Analysis: at user-month level (initially) and user-level (maybe later)
- 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)
- 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[1]

Old version - reuse what is useful.

- This paper explores the effect of spending profiles on financial outcomes.
- Spend profile: distribution of txn spend txns across txn categories.
- Financial outcomes: spending, saving, overdraft fees
- Financial wellbeing is important.
- People in UK and US also don't have enough to cover unexpected outlays. (See reports)
- This has important consequences:
  - Short-term: financial well-being (see reports)
  - Long-term (viscious cycle): scarcity hypothesis - makes it harder to focus on important things (plan for retirement, focus on healthy lifestyle, support children, ...) and might lead to vicious cycle (less savings leading to increased risk of financial hardship leading to more stress leading to less savings...)
- Spending and savings behaviour is important component - it's not all about lack of income.
- Structure paper:
  - Part I: describtive states of when people spend, save, and pay od fees
  - Part II: define measures that characterise spend profile

[1] 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.

  - Part III: regression analysis
  - Part IV: cluster analysis: do outcomes differ by groups? (poss sep ml focused paper)
  - Part V: predict non-saving, high-spend, and od fee months (poss in separate paper with above section)

- While there is large literature on long-term / pension savings, very little work on short-term financial spending and savings behaviour.
- In doing so, we aim to contribute to three literatures:
  - Main 1: Understanding emergency savings behaivour (nest, aspen reports)
  - 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)
-

2 Results

2.1 Main results

Table 1: has_sa_inflows results

Dependent Variable:	has_sa_inflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag)	0.040***				0.015***			
	[0.034; 0.046]				[0.005; 0.025]			
Entropy (auto-tag)		0.054***				0.032***		
		[0.048; 0.060]				[0.021; 0.043]		
Entropy (merchant)			0.061***				0.027***	
			[0.055; 0.068]				[0.016; 0.037]	
Entropy (grocery shop)				0.027***				0.010***
				[0.023; 0.030]				[0.005; 0.014]
Paid with credit (%)	-0.002***	-0.002***	-0.002***	-0.002***	7.99×10^{-5}	3.63×10^{-5}	-2.13×10^{-6}	1.41×10^{-5}
	[-0.002; -0.002]	[-0.002; -0.002]	[-0.002; -0.002]	[-0.002; -0.002]	[-0.0003; 0.0004]	[-0.0003; 0.0004]	[-0.0004; 0.0004]	[-0.0003; 0.0004]
Month spend	0.0009	-0.002	-0.003***	-0.0001	0.010***	0.009***	0.009***	0.010***
	[-0.001; 0.003]	[-0.004; 0.0005]	[-0.005; -0.001]	[-0.002; 0.002]	[0.008; 0.013]	[0.007; 0.012]	[0.007; 0.012]	[0.008; 0.013]
Urban	0.010*	0.011**	0.009*	0.009*	-0.027	-0.025	-0.024	-0.030
	[-1.3×10^{-5}; 0.019]	[0.002; 0.021]	[-0.0009; 0.018]	[-0.0008; 0.018]	[-0.235; 0.181]	[-0.237; 0.188]	[-0.236; 0.189]	[-0.237; 0.177]
Month income (Â£'000s)	-0.022***	-0.023***	-0.023***	-0.021***	0.004	0.004	0.004	0.004
	[-0.025; -0.020]	[-0.026; -0.021]	[-0.025; -0.020]	[-0.023; -0.018]	[-0.004; 0.012]	[-0.004; 0.011]	[-0.004; 0.011]	[-0.003; 0.012]
Has income in month	0.117***	0.115***	0.110***	0.112***	0.044***	0.042***	0.041***	0.044***
	[0.091; 0.143]	[0.089; 0.141]	[0.084; 0.136]	[0.086; 0.138]	[0.016; 0.072]	[0.014; 0.069]	[0.014; 0.069]	[0.016; 0.072]
income_variability	-0.021***	-0.022***	-0.021***	-0.020***	-0.003	-0.003	-0.003	-0.003
	[-0.026; -0.016]	[-0.027; -0.016]	[-0.026; -0.016]	[-0.025; -0.015]	[-0.013; 0.007]	[-0.013; 0.007]	[-0.013; 0.007]	[-0.013; 0.007]
has_rent_payment	-0.010*	-0.013***	-0.009*	-0.014***	0.028***	0.025**	0.028***	0.028***
	[-0.020; 0.0001]	[-0.024; -0.003]	[-0.019; 0.0007]	[-0.024; -0.004]	[0.009; 0.048]	[0.006; 0.045]	[0.008; 0.047]	[0.008; 0.048]
has_mortgage_payment	-0.020***	-0.028***	-0.024***	-0.018***	0.021*	0.016	0.018	0.021*
	[-0.029; -0.011]	[-0.036; -0.019]	[-0.033; -0.015]	[-0.027; -0.010]	[-0.004; 0.045]	[-0.009; 0.040]	[-0.006; 0.043]	[-0.004; 0.046]
Loan repayment	0.016***	0.011***	0.011***	0.020***	0.009	0.007	0.008	0.010
	[0.009; 0.024]	[0.003; 0.018]	[0.004; 0.018]	[0.013; 0.027]	[-0.007; 0.025]	[-0.009; 0.023]	[-0.008; 0.024]	[-0.006; 0.026]
Benefits	0.026***	0.023***	0.021***	0.019***	0.005	0.004	0.003	0.003
	[0.012; 0.039]	[0.009; 0.036]	[0.007; 0.034]	[0.006; 0.033]	[-0.033; 0.043]	[-0.034; 0.041]	[-0.035; 0.041]	[-0.035; 0.041]
(Intercept)	0.323***	0.329***	0.333***	0.337***				
	[0.295; 0.351]	[0.301; 0.357]	[0.305; 0.360]	[0.310; 0.365]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.01816	0.01959	0.02097	0.01902	0.42674	0.42711	0.42699	0.42680
Within R^2					0.00295	0.00360	0.00339	0.00305

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

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

Table 2: has_sa_inflows results

Dependent Variable:	has_sa_inflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag, smooth)	8.26×10^{-5} [-0.004; 0.004]				-0.003 [-0.009; 0.003]			
Entropy (auto-tag, smooth)		-0.031*** [-0.035; -0.027]				-0.019*** [-0.026; -0.011]		
Entropy (merchant, smooth)			-0.028*** [-0.032; -0.024]				-0.016*** [-0.024; -0.009]	
Entropy (grocery, smooth)				-0.018*** [-0.022; -0.015]				-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]	7.08×10^{-5} [-0.0003; 0.0004]	-7.13×10^{-5} [-0.0004; 0.0003]	-5.56×10^{-5} [-0.0004; 0.0003]	1.98×10^{-5} [-0.0003; 0.0004]
Month spend	0.002** [1.85×10^{-5}; 0.004]	-0.003** [-0.005; -0.0006]	-0.002** [-0.004; -5×10^{-5}]	0.0005 [-0.002; 0.003]	0.011*** [0.008; 0.013]	0.009*** [0.006; 0.011]	0.009*** [0.006; 0.011]	0.010*** [0.007; 0.013]
Urban	0.009* [-0.0005; 0.019]	0.003 [-0.006; 0.013]	0.005 [-0.005; 0.015]	0.008 [-0.002; 0.017]	-0.029 [-0.235; 0.177]	-0.027 [-0.233; 0.179]	-0.028 [-0.233; 0.178]	-0.031 [-0.237; 0.174]
Month income (Â£'000s)	-0.021*** [-0.024; -0.019]	-0.023*** [-0.026; -0.020]	-0.023*** [-0.026; -0.021]	-0.021*** [-0.024; -0.019]	0.004 [-0.003; 0.012]	0.003 [-0.005; 0.011]	0.003 [-0.005; 0.010]	0.004 [-0.004; 0.011]
Has income in month	0.119*** [0.092; 0.145]	0.113*** [0.087; 0.139]	0.115*** [0.089; 0.141]	0.115*** [0.088; 0.141]	0.045*** [0.017; 0.072]	0.042*** [0.015; 0.070]	0.043*** [0.016; 0.071]	0.043*** [0.016; 0.071]
income_variability	-0.022*** [-0.027; -0.017]	-0.023*** [-0.028; -0.018]	-0.023*** [-0.028; -0.018]	-0.021*** [-0.027; -0.016]	-0.003 [-0.013; 0.007]	-0.003 [-0.013; 0.006]	-0.003 [-0.013; 0.006]	-0.003 [-0.013; 0.006]
has_rent_payment	-0.013** [-0.023; -0.003]	-0.013** [-0.023; -0.003]	-0.013** [-0.023; -0.003]	-0.013*** [-0.023; -0.003]	0.028*** [0.008; 0.048]	0.029*** [0.009; 0.049]	0.029*** [0.009; 0.049]	0.028*** [0.008; 0.048]
has_mortgage_payment	-0.018*** [-0.027; -0.009]	-0.012*** [-0.021; -0.003]	-0.013*** [-0.022; -0.005]	-0.016*** [-0.025; -0.007]	0.021* [-0.003; 0.046]	0.022* [-0.003; 0.046]	0.021* [-0.003; 0.046]	0.021* [-0.003; 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]	0.010 [-0.006; 0.026]	0.010 [-0.006; 0.026]	0.010 [-0.006; 0.026]	0.010 [-0.006; 0.026]
Benefits	0.021*** [0.007; 0.034]	0.016** [0.002; 0.029]	0.016** [0.002; 0.029]	0.020*** [0.006; 0.033]	0.004 [-0.034; 0.042]	0.002 [-0.035; 0.040]	0.003 [-0.035; 0.040]	0.004 [-0.034; 0.042]
(Intercept)	0.331*** [0.303; 0.358]	0.348*** [0.320; 0.376]	0.348*** [0.320; 0.376]	0.337*** [0.310; 0.365]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.01636	0.01909	0.01844	0.01769	0.42662	0.42701	0.42693	0.42679
Within R^2					0.00275	0.00343	0.00328	0.00303

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

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

Table 3: sa_inflows results

Dependent Variable:	sa_inflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag)	26.9***				26.6***			
	[19.7; 34.0]				[16.4; 36.8]			
Entropy (auto-tag)		44.1***				46.5***		
		[37.0; 51.3]				[35.0; 57.9]		
Entropy (merchant)			42.6***				44.7***	
			[35.8; 49.4]				[33.5; 55.8]	
Entropy (grocery shop)				16.0***				9.88***
				[12.1; 19.9]				[4.52; 15.2]
Paid with credit (%)	-0.798***	-0.827***	-0.831***	-0.866***	-0.220	-0.283	-0.356*	-0.285
	[-0.985; -0.610]	[-1.01; -0.639]	[-1.02; -0.644]	[-1.05; -0.679]	[-0.587; 0.147]	[-0.650; 0.084]	[-0.723; 0.011]	[-0.653; 0.083]
Month spend	21.1***	18.8***	18.2***	20.5***	27.4***	26.1***	25.7***	27.3***
	[18.8; 23.3]	[16.5; 21.1]	[15.9; 20.6]	[18.3; 22.8]	[23.4; 31.4]	[22.1; 30.0]	[21.7; 29.6]	[23.3; 31.3]
Urban	7.82	9.21*	7.21	7.33	-79.1	-76.4	-74.3	-83.8
	[-2.73; 18.4]	[-1.34; 19.8]	[-3.34; 17.8]	[-3.22; 17.9]	[-324.2; 166.0]	[-327.5; 174.8]	[-326.9; 178.4]	[-328.2; 160.5]
Month income (Â£'000s)	0.718	-0.179	0.373	1.71	1.32	0.880	1.13	1.62
	[-2.28; 3.72]	[-3.18; 2.82]	[-2.62; 3.37]	[-1.29; 4.71]	[-8.02; 10.7]	[-8.44; 10.2]	[-8.18; 10.4]	[-7.72; 11.0]
Has income in month	89.5***	87.5***	84.7***	86.6***	54.8***	51.9***	50.7***	55.7***
	[60.6; 118.4]	[58.6; 116.4]	[55.8; 113.6]	[57.7; 115.5]	[25.4; 84.3]	[22.4; 81.3]	[21.2; 80.1]	[26.2; 85.2]
income_variability	12.2***	11.9***	12.3***	12.8***	9.21	9.08	9.23	9.15
	[6.36; 18.1]	[6.01; 17.8]	[6.47; 18.2]	[6.88; 18.7]	[-2.71; 21.1]	[-2.85; 21.0]	[-2.69; 21.1]	[-2.77; 21.1]
has_rent_payment	-6.65	-9.06	-6.16	-9.18	15.3	10.6	14.1	14.5
	[-17.8; 4.54]	[-20.2; 2.10]	[-17.3; 5.02]	[-20.3; 1.99]	[-4.93; 35.6]	[-9.52; 30.7]	[-6.12; 34.3]	[-5.76; 34.7]
has_mortgage_payment	-17.3***	-24.1***	-20.2***	-16.3***	0.921	-6.41	-3.17	1.56
	[-27.0; -7.67]	[-33.8; -14.3]	[-29.9; -10.6]	[-26.0; -6.68]	[-26.6; 28.4]	[-33.8; 21.0]	[-30.5; 24.2]	[-25.9; 29.0]
Loan repayment	-8.10**	-13.7***	-12.0***	-5.43	-6.07	-9.18	-8.34	-4.26
	[-16.0; -0.239]	[-21.6; -5.74]	[-19.9; -4.14]	[-13.2; 2.36]	[-22.5; 10.4]	[-25.6; 7.25]	[-24.8; 8.10]	[-20.7; 12.2]
Benefits	11.3	9.68	8.00	7.04	-5.57	-7.94	-9.47	-8.18
	[-3.76; 26.4]	[-5.38; 24.7]	[-7.06; 23.1]	[-8.02; 22.1]	[-45.5; 34.4]	[-47.9; 32.0]	[-49.5; 30.6]	[-48.3; 31.9]
(Intercept)	105.6***	109.1***	112.0***	114.6***				
	[74.8; 136.3]	[78.4; 139.8]	[81.3; 142.6]	[83.9; 145.3]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00812	0.00923	0.00928	0.00825	0.32189	0.32246	0.32244	0.32173
Within R^2					0.00814	0.00897	0.00894	0.00790

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

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

Table 4: sa_inflows results

Dependent Variable:	sa_inflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag, smooth)	0.895				3.72			
	[-3.56; 5.35]				[-2.83; 10.3]			
Entropy (auto-tag, smooth)		-20.0***				-16.3***		
		[-24.4; -15.5]				[-24.4; -8.29]		
Entropy (merchant, smooth)			-20.7***				-15.0***	
			[-25.3; -16.2]				[-23.3; -6.78]	
Entropy (grocery, smooth)				-12.1***				-6.85**
				[-15.9; -8.38]				[-13.5; -0.242]
Paid with credit (%)	-0.831***	-0.905***	-0.942***	-0.873***	-0.198	-0.348*	-0.342*	-0.257
	[-1.02; -0.643]	[-1.09; -0.717]	[-1.13; -0.754]	[-1.06; -0.686]	[-0.565; 0.168]	[-0.719; 0.022]	[-0.713; 0.029]	[-0.625; 0.111]
Month spend	21.9***	18.8***	18.7***	20.8***	28.0***	26.0***	26.1***	27.4***
	[19.6; 24.1]	[16.4; 21.1]	[16.3; 21.0]	[18.5; 23.1]	[24.0; 31.9]	[22.0; 30.0]	[22.0; 30.1]	[23.4; 31.4]
Urban	7.58	3.91	4.45	6.55	-82.5	-81.1	-81.8	-84.3
	[-2.98; 18.1]	[-6.67; 14.5]	[-6.12; 15.0]	[-4.01; 17.1]	[-325.5; 160.5]	[-323.5; 161.2]	[-324.0; 160.4]	[-326.9; 158.3]
Month income (Â£'000s)	1.30	0.213	-0.100	1.35	1.71	0.612	0.417	1.38
	[-1.70; 4.29]	[-2.79; 3.22]	[-3.11; 2.91]	[-1.65; 4.34]	[-7.64; 11.1]	[-8.73; 9.96]	[-8.93; 9.76]	[-7.98; 10.7]
Has income in month	90.6***	86.8***	87.9***	87.8***	56.9***	54.2***	55.2***	55.7***
	[61.7; 119.5]	[57.9; 115.7]	[59.0; 116.8]	[58.8; 116.7]	[27.3; 86.6]	[24.7; 83.7]	[25.7; 84.8]	[26.1; 85.3]
income_variability	11.8***	11.2***	11.1***	12.0***	9.08	8.76	8.73	8.89
	[5.94; 17.7]	[5.28; 17.0]	[5.17; 16.9]	[6.11; 17.9]	[-2.83; 21.0]	[-3.12; 20.6]	[-3.16; 20.6]	[-3.00; 20.8]
has_rent_payment	-8.68	-8.83	-8.77	-8.89	15.0	15.4	15.4	14.9
	[-19.9; 2.51]	[-20.0; 2.34]	[-19.9; 2.39]	[-20.1; 2.28]	[-5.30; 35.2]	[-4.87; 35.6]	[-4.78; 35.6]	[-5.36; 35.1]
has_mortgage_payment	-16.2***	-12.4**	-12.6**	-14.9***	1.70	2.13	1.98	1.81
	[-25.8; -6.48]	[-22.1; -2.72]	[-22.3; -2.90]	[-24.5; -5.18]	[-25.8; 29.2]	[-25.3; 29.5]	[-25.4; 29.4]	[-25.6; 29.2]
Loan repayment	-4.34	-2.97	-2.83	-3.45	-4.33	-3.89	-3.82	-4.02
	[-12.2; 3.49]	[-10.8; 4.83]	[-10.6; 4.96]	[-11.2; 4.35]	[-20.8; 12.1]	[-20.3; 12.5]	[-20.2; 12.6]	[-20.4; 12.4]
Benefits	8.06	4.84	4.18	7.34	-6.29	-8.64	-8.38	-6.91
	[-7.04; 23.2]	[-10.2; 19.9]	[-10.9; 19.3]	[-7.72; 22.4]	[-46.3; 33.7]	[-48.7; 31.4]	[-48.4; 31.7]	[-47.0; 33.2]
(Intercept)	110.5***	121.8***	123.5***	115.2***				
	[79.8; 141.1]	[91.0; 152.6]	[92.7; 154.3]	[84.4; 145.9]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00745	0.00837	0.00840	0.00794	0.32158	0.32182	0.32179	0.32163
Within R^2					0.00769	0.00803	0.00799	0.00776

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

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

Table 5: sa_netflows results

Dependent Variable:	sa_netflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag)	26.4***				20.1***			
	[19.4; 33.4]				[10.1; 30.0]			
Entropy (auto-tag)		36.6***				28.0***		
		[29.6; 43.6]				[17.1; 38.9]		
Entropy (merchant)			35.7***				31.1***	
			[29.1; 42.4]				[20.5; 41.8]	
Entropy (grocery shop)				4.96**				3.04
				[1.16; 8.75]				[-2.11; 8.19]
Paid with credit (%)	-0.918***	-0.947***	-0.951***	-0.964***	-0.718***	-0.755***	-0.812***	-0.735***
	[-1.10; -0.734]	[-1.13; -0.764]	[-1.13; -0.768]	[-1.15; -0.780]	[-1.08; -0.353]	[-1.12; -0.389]	[-1.18; -0.446]	[-1.10; -0.369]
Month spend	21.0***	19.2***	18.7***	21.3***	20.3***	19.6***	19.1***	20.5***
	[18.7; 23.2]	[17.0; 21.5]	[16.4; 21.0]	[19.1; 23.6]	[16.6; 24.1]	[15.8; 23.3]	[15.3; 22.9]	[16.7; 24.2]
Urban	-16.5***	-15.3***	-17.0***	-16.8***	-75.8	-74.6	-72.6	-78.8
	[-26.8; -6.13]	[-25.7; -5.02]	[-27.3; -6.69]	[-27.1; -6.49]	[-236.5; 85.0]	[-236.4; 87.1]	[-233.6; 88.4]	[-238.8; 81.2]
Month income (Â£'000s)	-20.7***	-21.4***	-20.9***	-20.0***	-13.7***	-13.9***	-13.8***	-13.4***
	[-23.7; -17.8]	[-24.3; -18.5]	[-23.9; -18.0]	[-23.0; -17.1]	[-23.5; -3.84]	[-23.7; -4.06]	[-23.6; -3.96]	[-23.2; -3.58]
Has income in month	-7.99	-9.50	-11.9	-8.24	-26.4	-27.9*	-29.2*	-25.3
	[-36.3; 20.3]	[-37.8; 18.8]	[-40.2; 16.4]	[-36.5; 20.1]	[-58.4; 5.51]	[-59.9; 4.05]	[-61.2; 2.76]	[-57.3; 6.62]
income_variability	-5.81**	-6.16**	-5.78**	-5.94**	-10.2*	-10.3*	-10.2*	-10.2*
	[-11.6; -0.047]	[-11.9; -0.404]	[-11.5; -0.025]	[-11.7; -0.173]	[-21.3; 0.986]	[-21.4; 0.865]	[-21.3; 0.979]	[-21.4; 0.894]
has_rent_payment	-14.0**	-16.3***	-13.9**	-16.2***	2.18	-0.730	1.29	1.71
	[-24.9; -3.05]	[-27.3; -5.40]	[-24.8; -2.96]	[-27.2; -5.29]	[-17.4; 21.8]	[-20.3; 18.8]	[-18.2; 20.8]	[-17.9; 21.3]
has_mortgage_payment	-19.9***	-25.3***	-22.2***	-18.7***	-23.6**	-27.9**	-26.4**	-23.1*
	[-29.4; -10.5]	[-34.8; -15.8]	[-31.6; -12.7]	[-28.2; -9.29]	[-47.1; -0.174]	[-51.4; -4.42]	[-49.9; -2.97]	[-46.5; 0.344]
Loan repayment	11.0***	6.99*	8.30**	14.5***	-7.54	-9.11	-9.01	-6.08
	[3.33; 18.7]	[-0.775; 14.8]	[0.578; 16.0]	[6.85; 22.1]	[-21.3; 6.27]	[-22.9; 4.70]	[-22.8; 4.79]	[-19.9; 7.69]
Benefits	10.7	8.81	7.41	7.05	4.14	2.54	1.36	2.82
	[-4.05; 25.5]	[-5.93; 23.5]	[-7.32; 22.1]	[-7.70; 21.8]	[-27.4; 35.7]	[-28.9; 34.0]	[-30.2; 32.9]	[-28.8; 34.4]
(Intercept)	9.97	13.7	16.1	16.1				
	[-20.1; 40.0]	[-16.3; 43.7]	[-13.9; 46.1]	[-13.9; 46.2]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00807	0.00868	0.00874	0.00748	0.13051	0.13066	0.13077	0.13034
Within R^2					0.00354	0.00371	0.00383	0.00334

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

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

Table 6: sa_netflows results

Dependent Variable:	sa_netflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag, smooth)	7.74***				10.1***			
	[3.38; 12.1]				[3.75; 16.5]			
Entropy (auto-tag, smooth)		-9.53***				0.983		
		[-13.9; -5.14]				[-6.90; 8.86]		
Entropy (merchant, smooth)			-4.88**				8.20*	
			[-9.38; -0.386]				[-0.347; 16.8]	
Entropy (grocery, smooth)				-2.33				4.15
				[-6.00; 1.35]				[-1.88; 10.2]
Paid with credit (%)	-0.926***	-0.988***	-0.979***	-0.961***	-0.671***	-0.705***	-0.644***	-0.687***
	[-1.11; -0.742]	[-1.17; -0.804]	[-1.16; -0.795]	[-1.15; -0.778]	[-1.04; -0.304]	[-1.07; -0.336]	[-1.02; -0.272]	[-1.06; -0.319]
Month spend	22.0***	20.3***	21.0***	21.5***	20.9***	20.7***	21.6***	20.9***
	[19.8; 24.2]	[18.0; 22.6]	[18.7; 23.3]	[19.3; 23.8]	[17.2; 24.7]	[16.9; 24.6]	[17.7; 25.5]	[17.1; 24.7]
Urban	-16.1***	-18.5***	-17.5***	-16.9***	-77.8	-78.6	-79.0	-77.5
	[-26.5; -5.79]	[-28.8; -8.12]	[-27.8; -7.13]	[-27.3; -6.61]	[-238.4; 82.7]	[-238.9; 81.8]	[-239.6; 81.6]	[-237.9; 82.9]
Month income (Â£'000s)	-20.2***	-20.7***	-20.5***	-20.2***	-13.3***	-13.3***	-12.7**	-13.2***
	[-23.2; -17.3]	[-23.6; -17.7]	[-23.4; -17.5]	[-23.1; -17.2]	[-23.2; -3.53]	[-23.2; -3.46]	[-22.5; -2.83]	[-23.0; -3.36]
Has income in month	-6.03	-8.78	-7.65	-7.56	-24.3	-24.9	-24.3	-24.4
	[-34.3; 22.3]	[-37.1; 19.5]	[-35.9; 20.6]	[-35.9; 20.7]	[-56.3; 7.68]	[-57.0; 7.23]	[-56.3; 7.81]	[-56.5; 7.61]
income_variability	-6.07**	-6.54**	-6.41**	-6.20**	-10.2*	-10.3*	-10.1*	-10.2*
	[-11.8; -0.307]	[-12.3; -0.781]	[-12.2; -0.649]	[-12.0; -0.439]	[-21.3; 0.977]	[-21.4; 0.868]	[-21.2; 1.00]	[-21.3; 0.940]
has_rent_payment	-15.2***	-16.1***	-16.1***	-16.1***	2.12	1.79	1.51	1.81
	[-26.2; -4.27]	[-27.1; -5.19]	[-27.0; -5.16]	[-27.1; -5.18]	[-17.4; 21.7]	[-17.8; 21.4]	[-18.1; 21.1]	[-17.8; 21.4]
has_mortgage_payment	-19.5***	-16.9***	-17.8***	-18.4***	-23.0*	-23.1*	-23.2*	-23.1*
	[-28.9; -10.0]	[-26.4; -7.43]	[-27.3; -8.35]	[-27.9; -8.96]	[-46.5; 0.387]	[-46.5; 0.367]	[-46.6; 0.269]	[-46.5; 0.345]
Loan repayment	13.5***	15.4***	15.2***	15.0***	-6.77	-6.03	-6.15	-6.04
	[5.85; 21.2]	[7.82; 23.1]	[7.55; 22.8]	[7.38; 22.6]	[-20.6; 7.01]	[-19.8; 7.74]	[-19.9; 7.61]	[-19.8; 7.73]
Benefits	8.95	5.86	6.44	7.20	4.47	3.38	4.16	3.37
	[-5.82; 23.7]	[-8.90; 20.6]	[-8.33; 21.2]	[-7.54; 21.9]	[-27.1; 36.0]	[-28.2; 34.9]	[-27.4; 35.7]	[-28.2; 34.9]
(Intercept)	14.2	20.3	17.9	15.8				
	[-15.9; 44.2]	[-9.87; 50.4]	[-12.2; 48.1]	[-14.3; 45.8]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00755	0.00762	0.00745	0.00742	0.13044	0.13032	0.13039	0.13035
Within R^2					0.00346	0.00332	0.00340	0.00335

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

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

Table 7: sa_outflows results

Dependent Variable:	sa_outflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag)	0.444				6.57			
	[-6.48; 7.36]				[-3.62; 16.8]			
Entropy (auto-tag)		7.53**				18.4***		
		[0.579; 14.5]				[6.40; 30.5]		
Entropy (merchant)			6.92**				13.5**	
			[0.307; 13.5]				[2.36; 24.7]	
Entropy (grocery shop)				11.0***				6.84**
				[7.28; 14.8]				[1.43; 12.2]
Paid with credit (%)	0.120	0.121	0.120	0.098	0.498**	0.472**	0.456**	0.450**
	[-0.062; 0.303]	[-0.061; 0.303]	[-0.062; 0.302]	[-0.085; 0.280]	[0.118; 0.878]	[0.092; 0.852]	[0.076; 0.837]	[0.072; 0.829]
Month spend	0.092	-0.410	-0.482	-0.799	7.11***	6.51***	6.56***	6.85***
	[-2.12; 2.30]	[-2.66; 1.84]	[-2.75; 1.79]	[-3.02; 1.42]	[4.11; 10.1]	[3.52; 9.50]	[3.57; 9.55]	[3.85; 9.85]
Urban	24.3***	24.6***	24.2***	24.1***	-3.34	-1.71	-1.67	-5.01
	[14.0; 34.5]	[14.3; 34.8]	[14.0; 34.5]	[13.9; 34.4]	[-276.0; 269.3]	[-276.5; 273.1]	[-276.6; 273.2]	[-278.5; 268.5]
Month income (Â£'000s)	21.5***	21.2***	21.3***	21.8***	15.0***	14.8***	14.9***	15.0***
	[18.5; 24.4]	[18.3; 24.1]	[18.4; 24.2]	[18.8; 24.7]	[4.69; 25.3]	[4.46; 25.0]	[4.60; 25.2]	[4.72; 25.3]
Has income in month	97.5***	97.0***	96.6***	94.8***	81.3***	79.8***	79.9***	81.0***
	[69.4; 125.6]	[68.9; 125.1]	[68.4; 124.7]	[66.7; 122.9]	[52.1; 110.4]	[50.7; 108.9]	[50.9; 109.0]	[52.0; 110.1]
income_variability	18.0***	18.1***	18.1***	18.7***	19.4***	19.3***	19.4***	19.4***
	[12.3; 23.8]	[12.3; 23.8]	[12.4; 23.9]	[13.0; 24.4]	[6.78; 32.0]	[6.74; 31.9]	[6.79; 32.0]	[6.82; 32.0]
has_rent_payment	7.35	7.27	7.74	7.04	13.1	11.3	12.8	12.8
	[-3.52; 18.2]	[-3.59; 18.1]	[-3.13; 18.6]	[-3.82; 17.9]	[-7.32; 33.6]	[-9.14; 31.8]	[-7.68; 33.2]	[-7.69; 33.2]
has_mortgage_payment	2.57	1.23	1.91	2.40	24.5	21.5	23.3	24.6
	[-6.83; 12.0]	[-8.25; 10.7]	[-7.50; 11.3]	[-6.99; 11.8]	[-5.20; 54.3]	[-8.06; 51.1]	[-6.42; 52.9]	[-5.10; 54.4]
Loan repayment	-19.1***	-20.7***	-20.3***	-19.9***	1.47	-0.070	0.667	1.82
	[-26.8; -11.5]	[-28.4; -12.9]	[-28.0; -12.6]	[-27.5; -12.3]	[-15.4; 18.4]	[-16.9; 16.8]	[-16.2; 17.5]	[-15.0; 18.7]
Benefits	0.622	0.873	0.585	-0.004	-9.71	-10.5	-10.8	-11.0
	[-14.1; 15.3]	[-13.8; 15.5]	[-14.1; 15.2]	[-14.6; 14.6]	[-47.4; 28.0]	[-48.2; 27.3]	[-48.6; 26.9]	[-48.8; 26.8]
(Intercept)	95.6***	95.4***	95.9***	98.5***				
	[65.7; 125.4]	[65.6; 125.3]	[66.1; 125.7]	[68.6; 128.3]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00669	0.00674	0.00674	0.00709	0.38593	0.38606	0.38600	0.38599
Within R^2					0.00292	0.00313	0.00303	0.00302

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

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

Table 8: sa_outflows results

Dependent Variable:	sa_outflows							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Entropy (tag, smooth)	-6.85***				-6.39*			
	[-11.2; -2.51]				[-13.0; 0.242]			
Entropy (auto-tag, smooth)		-10.4***				-17.3***		
		[-14.8; -6.06]				[-25.6; -8.98]		
Entropy (merchant, smooth)			-15.9***				-23.2***	
			[-20.3; -11.4]				[-31.8; -14.7]	
Entropy (grocery, smooth)				-9.81***				-11.0***
				[-13.5; -6.16]				[-17.9; -4.07]
Paid with credit (%)	0.095	0.083	0.037	0.088	0.472**	0.357*	0.302	0.430**
	[-0.088; 0.278]	[-0.100; 0.266]	[-0.147; 0.221]	[-0.094; 0.271]	[0.092; 0.853]	[-0.028; 0.742]	[-0.085; 0.689]	[0.047; 0.813]
Month spend	-0.093	-1.51	-2.32**	-0.761	7.02***	5.24***	4.46***	6.45***
	[-2.30; 2.11]	[-3.81; 0.794]	[-4.62; -0.019]	[-2.98; 1.46]	[4.01; 10.0]	[2.18; 8.29]	[1.40; 7.52]	[3.45; 9.44]
Urban	23.7***	22.4***	21.9***	23.5***	-4.63	-2.57	-2.79	-6.78
	[13.4; 34.0]	[12.1; 32.7]	[11.7; 32.2]	[13.2; 33.8]	[-276.7; 267.4]	[-274.7; 269.6]	[-274.5; 269.0]	[-278.7; 265.1]
Month income (Â£'000s)	21.5***	20.9***	20.4***	21.5***	15.1***	13.9***	13.1**	14.6***
	[18.6; 24.5]	[18.0; 23.8]	[17.5; 23.3]	[18.6; 24.4]	[4.76; 25.4]	[3.62; 24.2]	[2.82; 23.4]	[4.26; 24.9]
Has income in month	96.6***	95.6***	95.5***	95.3***	81.3***	79.1***	79.5***	80.1***
	[68.5; 124.7]	[67.5; 123.7]	[67.4; 123.6]	[67.2; 123.4]	[52.1; 110.4]	[49.9; 108.2]	[50.4; 108.5]	[51.0; 109.3]
income_variability	17.9***	17.7***	17.5***	18.2***	19.2***	19.0***	18.8***	19.1***
	[12.2; 23.6]	[12.0; 23.4]	[11.7; 23.2]	[12.5; 23.9]	[6.67; 31.8]	[6.50; 31.6]	[6.33; 31.4]	[6.55; 31.6]
has_rent_payment	6.53	7.29	7.32	7.23	12.8	13.6	13.9	13.1
	[-4.34; 17.4]	[-3.57; 18.1]	[-3.54; 18.2]	[-3.63; 18.1]	[-7.62; 33.3]	[-6.86; 34.0]	[-6.48; 34.3]	[-7.39; 33.5]
has_mortgage_payment	3.31	4.50	5.25	3.57	24.7	25.2*	25.2*	24.9
	[-6.09; 12.7]	[-4.93; 13.9]	[-4.17; 14.7]	[-5.83; 13.0]	[-5.01; 54.5]	[-4.55; 54.9]	[-4.61; 54.9]	[-4.85; 54.6]
Loan repayment	-17.9***	-18.4***	-18.0***	-18.5***	2.44	2.14	2.33	2.02
	[-25.5; -10.2]	[-26.0; -10.8]	[-25.6; -10.4]	[-26.0; -10.9]	[-14.4; 19.3]	[-14.7; 19.0]	[-14.5; 19.2]	[-14.8; 18.9]
Benefits	-0.892	-1.02	-2.25	0.140	-10.8	-12.0	-12.5	-10.3
	[-15.6; 13.8]	[-15.7; 13.6]	[-16.9; 12.4]	[-14.5; 14.8]	[-48.5; 27.0]	[-49.7; 25.7]	[-50.3; 25.2]	[-48.1; 27.6]
(Intercept)	96.3***	101.6***	105.6***	99.4***				
	[66.5; 126.1]	[71.6; 131.5]	[75.6; 135.5]	[69.5; 129.3]				
Fixed-effects								
User id					Yes	Yes	Yes	Yes
Calendar month					Yes	Yes	Yes	Yes
Fit statistics								
Observations	81,701	81,701	81,701	81,701	81,701	81,701	81,701	81,701
R^2	0.00680	0.00695	0.00728	0.00703	0.38596	0.38621	0.38648	0.38610
Within R^2					0.00297	0.00337	0.00380	0.00318

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

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

2.2 Exploration

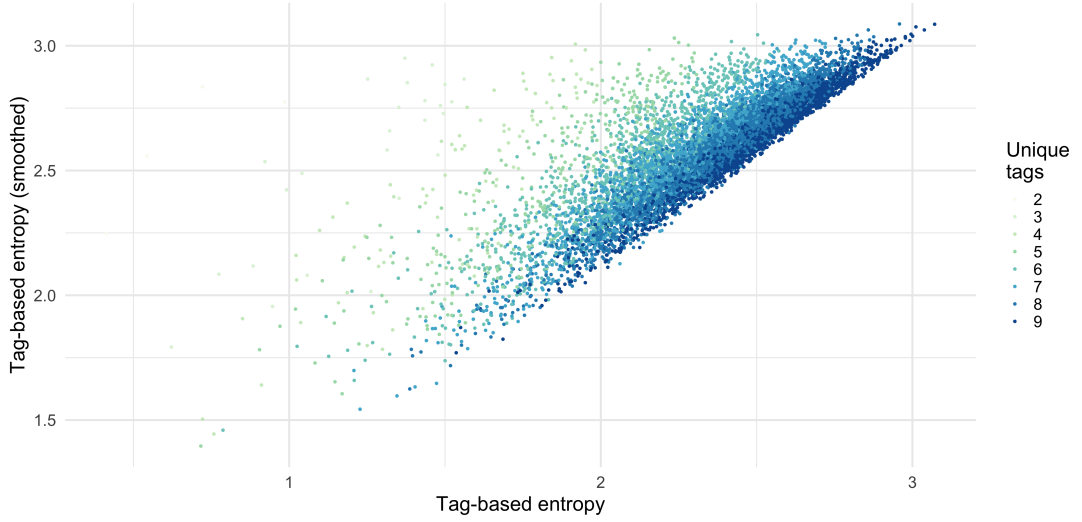
Table 9: sa_outflows results

Dependent Variable:	has_sa_inflows			
Model:	(1)	(2)	(3)	(4)
Variables				
Entropy (tag)	0.017***	0.0005	0.022***	-0.007
	[0.007; 0.026]	[-0.011; 0.012]	[0.012; 0.033]	[-0.020; 0.006]
Current account txns	0.001***			0.002***
	[0.001; 0.002]			[0.002; 0.002]
Paid with credit (%)	0.0003	-0.0003*	-0.0003*	0.0006***
	[-7.76×10^{-5}; 0.0006]	[-0.0007; 1.97×10^{-5}]	[-0.0007; 4.48×10^{-5}]	[0.0002; 0.0010]
Month spend	0.013***	0.019***	0.019***	0.015***
	[0.011; 0.016]	[0.017; 0.022]	[0.017; 0.022]	[0.012; 0.018]
Urban	-0.018	-0.023	-0.026	-0.018
	[-0.217; 0.181]	[-0.222; 0.176]	[-0.223; 0.172]	[-0.215; 0.180]
Month income (Â£'000s)	-0.0002	0.005	0.005	0.0006
	[-0.008; 0.007]	[-0.002; 0.013]	[-0.003; 0.013]	[-0.007; 0.008]
Has income in month	0.027*	0.038***	0.040***	0.026*
	[-0.002; 0.056]	[0.009; 0.067]	[0.011; 0.069]	[-0.003; 0.054]
income_variability	0.0009	0.003	0.002	0.001
	[-0.009; 0.011]	[-0.007; 0.012]	[-0.007; 0.012]	[-0.009; 0.011]
has_rent_payment	0.029***	0.028***	0.029***	0.028***
	[0.010; 0.048]	[0.009; 0.047]	[0.009; 0.048]	[0.009; 0.048]
has_mortgage_payment	0.005	0.005	0.006	0.006
	[-0.019; 0.030]	[-0.020; 0.030]	[-0.019; 0.031]	[-0.019; 0.031]
Loan repayment	0.001	0.004	0.004	0.0009
	[-0.015; 0.018]	[-0.013; 0.020]	[-0.012; 0.020]	[-0.015; 0.017]
Benefits	-0.005	0.005	0.005	-0.005
	[-0.041; 0.032]	[-0.032; 0.042]	[-0.032; 0.042]	[-0.042; 0.031]
Unique spend categories		0.013***		0.005*
		[0.008; 0.018]		[-0.0008; 0.010]
ssd_tag			8.53×10^{-6}***	-2.56×10^{-5}***
			[2.79×10^{-6}; 1.43×10^{-5}]	[-3.27×10^{-5}; -1.84×10^{-5}]
Fixed-effects				
User id	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes
Fit statistics				
Observations	89,169	89,169	89,169	89,169
R^2	0.42115	0.41683	0.41667	0.42200
Within R^2	0.01607	0.00873	0.00845	0.01751

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.

Figure 1: Entropy and unique tags



3 Discussion

4 Methods

4.1 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.[2] The dataset contains more than 500 million transactions made between 2012 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.[3]

4.2 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 10 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 10: 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

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

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

4.3 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).[4]

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.

4.4 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.[5]

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:[6]

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

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

[4]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.

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

[6]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.

We normalise H by $log(N_{SC})$, the entropy of completely random shopping behaviour, so that it takes value between 0 and 1.[7]

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

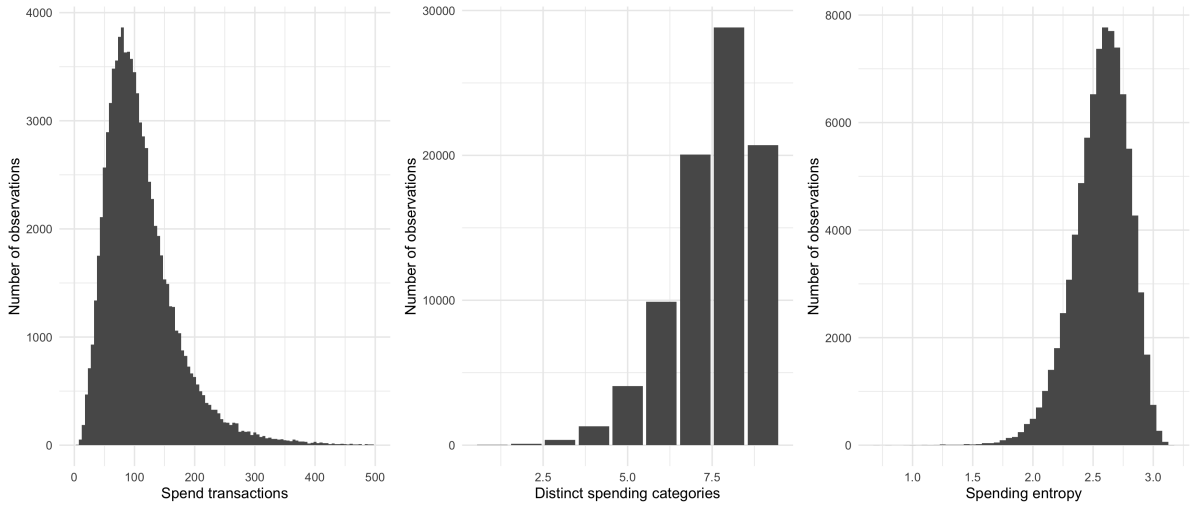
To calculate entropy scores, we group spending into 9 spending categories (SC) based on the classification used by Lloyds Banking Group as discussed in Muggleton et al. (2020). Also following that paper, we use additive smoothing to calcualte probabilities to avoid taking logs of zeroes in cases where 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.

Figure 2 shows the distribution of spending entropy as well as the distributions of the number of spend transactions, the number of distinct categories within which these transactions fall, and the proportion of transactions in each category.

Figure 2: Transactions distributions



Notes: From top-left to bottom-right: distribution of spending transactions per user-month, breakdown of spending transactions into spending categories; breakdown of number of spending categories spent on in user-month; distribution of user-month entropy scores.

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 ...

[7] $log(N_{SC})$ 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 $H = -N_{SC}p_i log(p_i) = -log(p_i) = log(N_{SC})$.

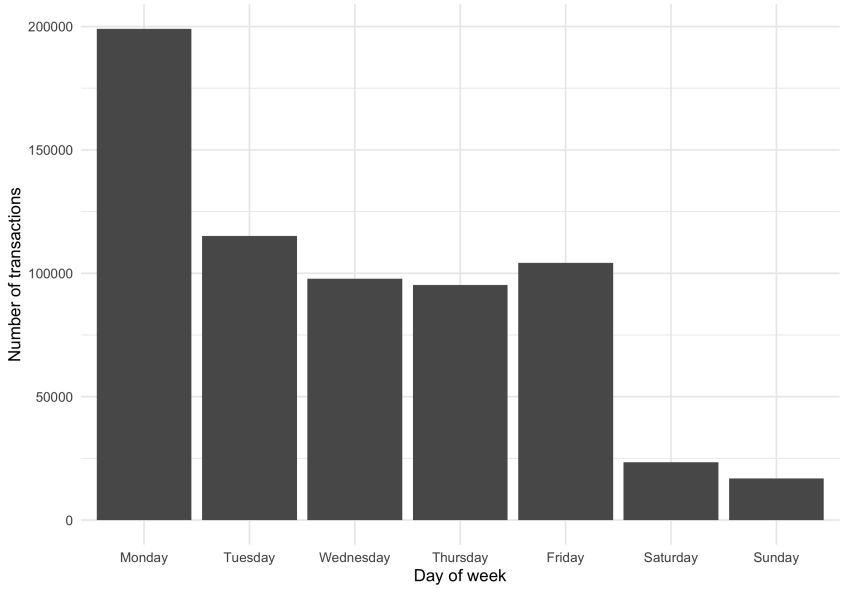
- All zero count-cases: ... Solution: require minimum number of txns in two different labels (for all categories we use to calculate entropy). Two to avoid 0 entropy cases that are unlikely to be genuine but probably artefacts of missing labelling.

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'.
- 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 shows in Figure 3, we cannot distinguish transactions made on Saturdays or Sundays from those made on Mondays, and thus classify all of them as weekend transactions.

Figure 3: Transactions by day of week



Notes: Number of transactions by day of week, based on a 1/1000 sample of the full data. Shows that banks process most weekend transactions on Mondays.

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.

4.5 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.[8]
- 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 –

[8] 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.

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 –

4.6 Summary statistics

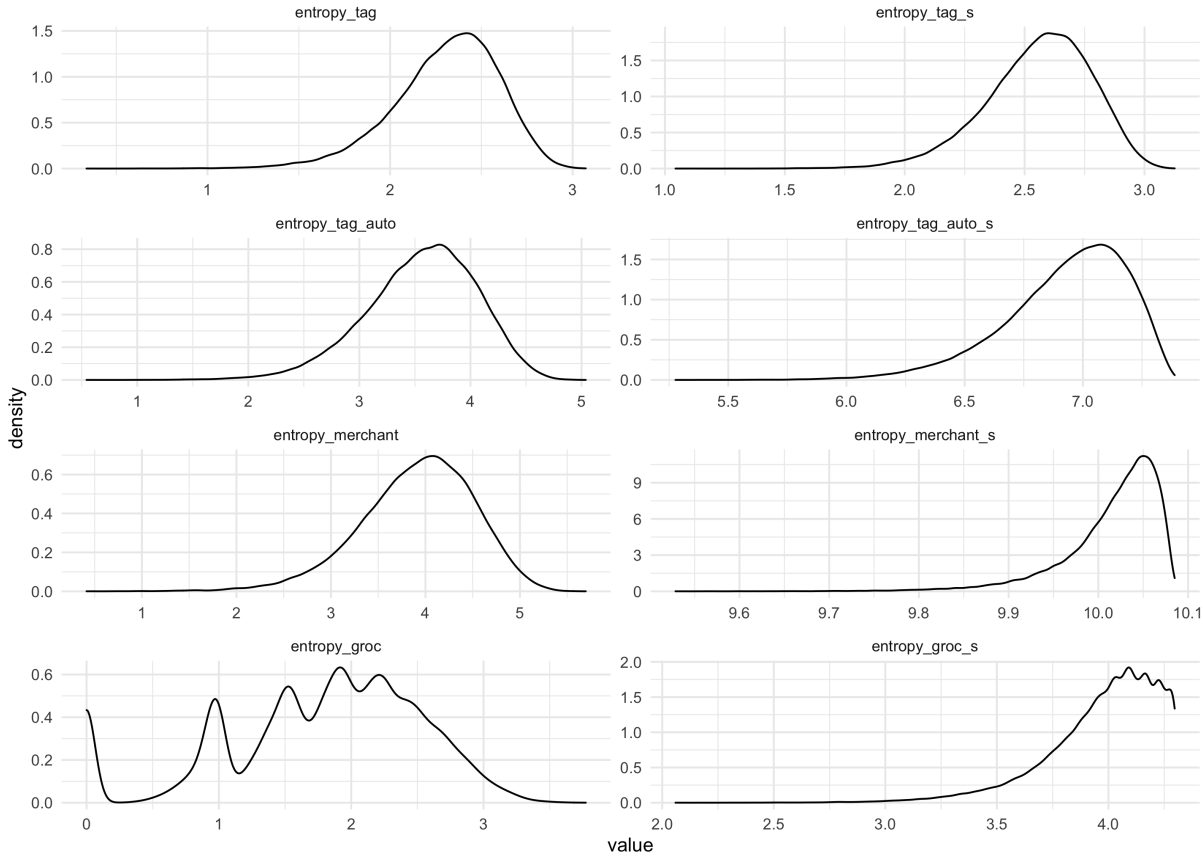
Table 11 provides summary statistics.

Figure 4

Table 11: Summary statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
user_id	337,760.700	176,496.800	37	190,247	383,977	480,137	589,517
ym	574.769	19.186	506	564	578	589	606
txns_count	111.040	50.989	6	74	101	139	294
txns_volume	15,882.750	17,854.630	256.890	6,069.430	10,102.470	18,037.340	150,708.400
txn_count_sa	4.277	7.905	0	0	2	5	167
txn_count_ca	91.117	46.488	5	57	83	117	293
nunique_tag	7.648	1.143	2	7	8	9	9
nunique_tag_auto	21.228	6.683	2	16	21	26	51
nunique_merchant	23.741	9.023	2	17	23	29	73
prop_credit	0.109	0.205	0.000	0.000	0.000	0.123	1.000
month_income_effective	2.738	1.793	0.000	1.577	2.268	3.438	9.249
month_income	2.719	1.535	0.833	1.616	2.290	3.373	9.249
year_income	32.627	18.416	10.000	19.398	27.480	40.472	110.991
income_variability	0.681	0.712	0.000	0.160	0.433	0.982	6.540
has_month_income	0.983	0.128	0	1	1	1	1
age	36.739	11.136	13	28	34	43	91
is_female	0.374	0.484	0	0	0	1	1
sa_inflows	632.739	1,721.472	0	0	0	403	11,430
sa_outflows	629.545	1,693.639	0	0	15	465.8	11,500
sa_netflows	3.194	1,744.711	−11,500	−100	0	0	11,430
has_sa_inflows	0.417	0.493	0	0	0	1	1
has_reg_sa_inflows	0.152	0.359	0	0	0	0	1
has_benefits	0.062	0.241	0	0	0	0	1
has_pension	0.031	0.174	0	0	0	0	1
has_rent_payment	0.127	0.333	0	0	0	0	1
has_mortgage_payment	0.227	0.419	0	0	0	0	1
has_loan_funds	0.030	0.169	0	0	0	0	1
has_loan_repmt	0.362	0.481	0	0	0	1	1
is_urban	0.860	0.347	0	1	1	1	1
spend_services	0.446	0.445	0.000	0.168	0.316	0.559	2.835
spend_other_spend	0.272	0.353	0.000	0.059	0.160	0.346	2.246
spend_household	0.978	1.003	0.0004	0.311	0.680	1.283	6.104
spend_travel	0.139	0.300	0	0	0.03	0.1	2
spend_communication	0.066	0.057	0	0.03	0.1	0.1	0
spend_finance	0.307	0.501	0	0.03	0.1	0.4	3
spend_motor	0.084	0.098	0	0	0.1	0.1	0
spend_retail	0.152	0.205	0	0.02	0.1	0.2	1
spend_hobbies	0.026	0.053	0	0	0.005	0.03	0
month_spend	2.613	2.208	0.201	1.246	1.979	3.223	14.954
has_od_fees	0.217	0.412	0	0	0	0	1
entropy_tag	2.306	0.290	0.337	2.138	2.340	2.513	3.071
entropy_tag_z	0.368	0.525	−3.194	0.063	0.428	0.741	1.752
entropy_tag_s	2.558	0.220	1.044	2.424	2.580	2.715	3.126
entropy_tag_sz	−0.144	0.827	−5.820	−0.645	−0.062	0.447	1.986
entropy_tag_auto	3.556	0.499	0.544	3.249	3.599	3.913	5.038
entropy_tag_auto_z	0.449	0.566	−2.965	0.101	0.497	0.853	2.128
entropy_tag_auto_s	6.926	0.262	5.277	6.779	6.970	7.121	7.388
entropy_tag_auto_sz	−0.316	0.894	−5.953	−0.820	−0.168	0.349	1.263
entropy_merchant	3.910	0.601	0.414	3.545	3.964	4.335	5.696
entropy_merchant_z	0.455	0.587	−2.959	0.099	0.508	0.870	2.200
entropy_merchant_s	10.015	0.055	9.529	9.993	10.029	10.053	10.085
entropy_merchant_sz	−0.227	0.896	−8.157	−0.578	0.003	0.392	0.920
entropy_groc	1.774	0.761	0	1.4	1.9	2.3	4
entropy_groc_z	0.151	0.956	−2.076	−0.355	0.305	0.839	2.659
entropy_groc_s	3.977	0.253	2.059	3.851	4.028	4.164	4.297
entropy_groc_sz	−0.104	0.986	−7.566	−0.592	0.095	0.624	1.142

Figure 4: Entropy distributions



Notes:

4.7 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?

References

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/.

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.

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.

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.

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.

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.

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:

A Data

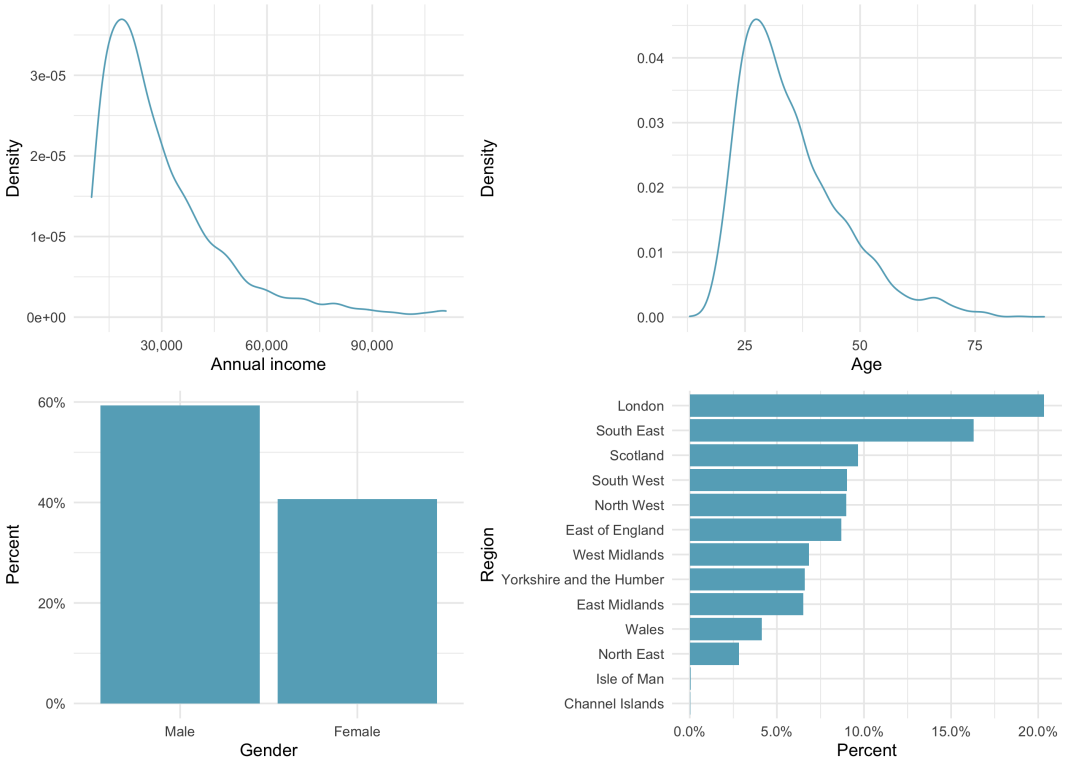
A.1 Data preprocessing

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 5: 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 12 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

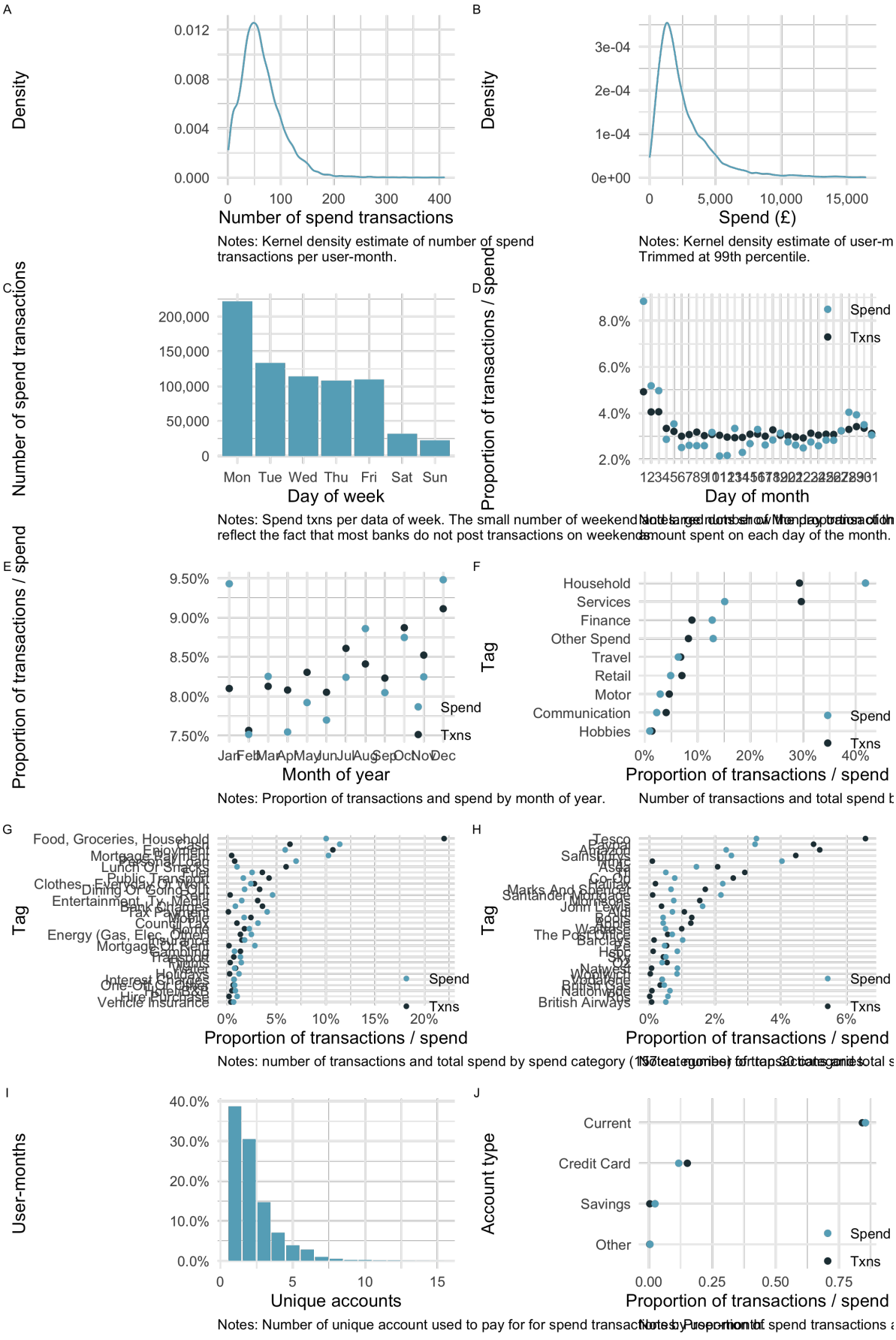
Figure 6

Table 12: 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 6: Spending behaviour



Notes: Kernel density estimate of number of spend transactions per user-month.

Notes: Kernel density estimate of user-m
Trimmed at 99th percentile.

Notes: Spend txns per data of week. The small number of weekend
reflect the fact that most banks do not post transactions on weekends

Notes: Proportion of transactions and spend by month of year.

Number of transactions and total spend b

Notes: number of transactions and total spend by spend category (

Notes: Number of unique account used to pay for for spend transactions

Figure 7: Savings behaviour