Entropy[*]

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

March 22, 2022

Abstract

...

Contents

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 New results

2.2 Effect of entropy on saving

Entropy measures are calculated based on the probability of occurrence of a pattern within a user month. Implicitly, this assumes that move diverse spending reflects more chaotic behaviour. The outcome variable is a dummy indicating whether a user made at least one transfer into their savings account in a given month.

- Table 17 shows results with a full set of controls for different FE specifications. Table 18 summarises shows controls added cumulatively.

- For smoothed entropy, the results are as hypothesised: higher entropy is strongly and robustly associated with a lower probability of making a savings transaction.

- Curiously, though, unsmoothed entropy is even more strongly but positively related to making a savings transaction.

- Figure 1 shows that smoothing increases entropy scores for those profiles with txns in few unique categories (the reason is that a low number of unique categories implies a high number of categories with zero counts, and smoothing reverses the implication of zero counts from 'zero entropy' to 'max entropy').

Table 1: Main results

Dependent Variable:			has_sa_inflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag	0.0495***					
	(0.0134)					
entropy_tag_n		0.1568***				
		(0.0425)				
Entropy (tag-based)			0.0273***			
			(0.0074)			
entropy_tag_s				0.0208		
				(0.0175)		
entropy_tag_sn					0.0660	
					(0.0556)	
Entropy (tag-based, smooth)						0.0056
						(0.0047)
Regular savings	0.1890***	0.1890***	0.1890***	0.1894***	0.1894***	0.1894***
	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)
Credit spend	-0.0035	-0.0035	-0.0035	-0.0015	-0.0015	-0.0015
	(0.0280)	(0.0280)	(0.0280)	(0.0281)	(0.0281)	(0.0281)
Month spend	0.0220***	0.0220***	0.0220***	0.0224***	0.0224***	0.0224***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Urban	0.1862*	0.1862*	0.1862*	0.1840*	0.1840*	0.1840*
	(0.1074)	(0.1074)	(0.1074)	(0.1040)	(0.1040)	(0.1040)
Month income ('000)	0.0023	0.0023	0.0023	0.0026	0.0026	0.0026
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)
Regular income	-0.0234**	-0.0234**	-0.0234**	-0.0235**	-0.0235**	-0.0235**
	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)
Has income in month	0.0206	0.0206	0.0206	0.0227	0.0227	0.0227
	(0.0242)	(0.0242)	(0.0242)	(0.0242)	(0.0242)	(0.0242)
Loan repayment	-0.0088	-0.0088	-0.0088	-0.0075	-0.0075	-0.0075
	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.41885	0.41885	0.41885	0.41854	0.41854	0.41854
Within R^2	0.02398	0.02398	0.02398	0.02346	0.02346	0.02346

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2: Main results

Dependent Variable:	has_sa_inflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag_auto	0.0470***					
	(0.0090)					
entropy_tag_auto_n		0.3479***				
		(0.0663)				
Entropy (auto-tag-based)			0.0415***			
			(0.0079)			
entropy_tag_auto_s				-0.0406**		
				(0.0179)		
entropy_tag_auto_sn					-0.3007**	
					(0.1324)	
Entropy (auto-tag-based, smooth)						-0.0119**
						(0.0052)
Regular savings	0.1888***	0.1888***	0.1888***	0.1894***	0.1894***	0.1894***
	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)
Credit spend	-0.0073	-0.0073	-0.0073	-0.0117	-0.0117	-0.0117
	(0.0280)	(0.0280)	(0.0280)	(0.0284)	(0.0284)	(0.0284)
Month spend	0.0210***	0.0210***	0.0210***	0.0210***	0.0210***	0.0210***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Urban	0.1879*	0.1879*	0.1879*	0.1837*	0.1837*	0.1837*
	(0.1112)	(0.1112)	(0.1112)	(0.1039)	(0.1039)	(0.1039)
Month income ('000)	0.0020	0.0020	0.0020	0.0018	0.0018	0.0018
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)
Regular income	-0.0240**	-0.0240**	-0.0240**	-0.0236**	-0.0236**	-0.0236**
	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)
Has income in month	0.0203	0.0203	0.0203	0.0223	0.0223	0.0223
	(0.0242)	(0.0242)	(0.0242)	(0.0242)	(0.0242)	(0.0242)
Loan repayment	-0.0108	-0.0108	-0.0108	-0.0070	-0.0070	-0.0070
	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.41926	0.41926	0.41926	0.41866	0.41866	0.41866
Within R^2	0.02467	0.02467	0.02467	0.02367	0.02367	0.02367

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3: Main results

Dependent Variable:			has_sa_inflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_merchant	0.0288***					
	(0.0076)					
entropy_merchant_n		0.2902***				
		(0.0764)				
Entropy (merchant-based)			0.0295***			
			(0.0078)			
entropy_merchant_s				-0.1107		
				(0.0750)		
entropy_merchant_sn					-1.117	
					(0.7567)	
Entropy (merchant-based, smooth)						-0.0068
						(0.0046)
Regular savings	0.1889***	0.1889***	0.1889***	0.1895***	0.1895***	0.1895***
	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)
Credit spend	-0.0085	-0.0085	-0.0085	-0.0088	-0.0088	-0.0088
	(0.0281)	(0.0281)	(0.0281)	(0.0285)	(0.0285)	(0.0285)
Month spend	0.0213***	0.0213***	0.0213***	0.0214***	0.0214***	0.0214***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Urban	0.1889*	0.1889*	0.1889*	0.1833*	0.1833*	0.1833*
	(0.1094)	(0.1094)	(0.1094)	(0.1037)	(0.1037)	(0.1037)
Month income ('000)	0.0022	0.0022	0.0022	0.0020	0.0020	0.0020
	(0.0027)	(0.0027)	(0.0027)	(0.0026)	(0.0026)	(0.0026)
Regular income	-0.0237**	-0.0237**	-0.0237**	-0.0237**	-0.0237**	-0.0237**
	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)
Has income in month	0.0203	0.0203	0.0203	0.0227	0.0227	0.0227
	(0.0241)	(0.0241)	(0.0241)	(0.0242)	(0.0242)	(0.0242)
Loan repayment	-0.0094	-0.0094	-0.0094	-0.0070	-0.0070	-0.0070
	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.41889	0.41889	0.41889	0.41857	0.41857	0.41857
Within R^2	0.02404	0.02404	0.02404	0.02352	0.02352	0.02352

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Main results

Dependent Variable:			has_sa_inflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_groc	0.0091*					
	(0.0052)					
entropy_groc_n		0.0301*				
		(0.0171)				
entropy_groc_z			0.0062*			
			(0.0035)			
entropy_groc_s				-0.0077		
				(0.0116)		
entropy_groc_sn					-0.0256	
					(0.0386)	
entropy_groc_sz						-0.0027
						(0.0041)
Regular savings	0.1893***	0.1893***	0.1893***	0.1894***	0.1894***	0.1894***
	(0.0110)	(0.0110)	(0.0110)	(0.0109)	(0.0109)	(0.0109)
Credit spend	-0.0040	-0.0040	-0.0040	-0.0040	-0.0040	-0.0040
	(0.0281)	(0.0281)	(0.0281)	(0.0283)	(0.0283)	(0.0283)
Month spend	0.0222***	0.0222***	0.0222***	0.0222***	0.0222***	0.0222***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Urban	0.1838*	0.1838*	0.1838*	0.1828*	0.1828*	0.1828*
	(0.1038)	(0.1038)	(0.1038)	(0.1036)	(0.1036)	(0.1036)
Month income ('000)	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025
	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)	(0.0026)
Regular income	-0.0236**	-0.0236**	-0.0236**	-0.0237**	-0.0237**	-0.0237**
	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)	(0.0099)
Has income in month	0.0224	0.0224	0.0224	0.0225	0.0225	0.0225
	(0.0241)	(0.0241)	(0.0241)	(0.0242)	(0.0242)	(0.0242)
Loan repayment	-0.0071	-0.0071	-0.0071	-0.0071	-0.0071	-0.0071
	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)	(0.0111)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.41856	0.41856	0.41856	0.41851	0.41851	0.41851
Within R^2	0.02350	0.02350	0.02350	0.02341	0.02341	0.02341

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 5: Main results

Dependent Variable:	sa_inflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag	62.68					
	(46.35)					
entropy_tag_n		198.7				
		(146.9)				
Entropy (tag-based)			34.65			
			(25.62)			
entropy_tag_s				114.4*		
				(62.59)		
entropy_tag_sn					362.6*	
					(198.4)	
Entropy (tag-based, smooth)						30.51*
						(16.69)
Regular savings	273.8***	273.8***	273.8***	273.9***	273.9***	273.9***
	(51.74)	(51.74)	(51.74)	(51.75)	(51.75)	(51.75)
Credit spend	-191.4*	-191.4*	-191.4*	-183.2*	-183.2*	-183.2*
	(110.2)	(110.2)	(110.2)	(109.9)	(109.9)	(109.9)
Month spend	235.3***	235.3***	235.3***	236.2***	236.2***	236.2***
	(16.37)	(16.37)	(16.37)	(16.37)	(16.37)	(16.37)
Urban	421.3	421.3	421.3	420.4	420.4	420.4
	(274.4)	(274.4)	(274.4)	(273.5)	(273.5)	(273.5)
Month income ('000)	-14.13	-14.13	-14.13	-13.69	-13.69	-13.69
	(14.82)	(14.82)	(14.82)	(14.83)	(14.83)	(14.83)
Regular income	-62.05**	-62.05**	-62.05**	-61.76**	-61.76**	-61.76**
	(30.85)	(30.85)	(30.85)	(30.84)	(30.84)	(30.84)
Has income in month	-10.06	-10.06	-10.06	-7.704	-7.704	-7.704
	(88.39)	(88.39)	(88.39)	(88.37)	(88.37)	(88.37)
Loan repayment	-117.2***	-117.2***	-117.2***	-117.1***	-117.1***	-117.1***
	(34.06)	(34.06)	(34.06)	(34.03)	(34.03)	(34.03)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.29119	0.29119	0.29119	0.29124	0.29124	0.29124
Within R^2	0.05470	0.05470	0.05470	0.05476	0.05476	0.05476

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 6: Main results

Dependent Variable:			sa_inflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag_auto	96.85***					
	(31.56)					
entropy_tag_auto_n		716.8***				
		(233.6)				
Entropy (auto-tag-based)			85.46***			
			(27.85)			
entropy_tag_auto_s				268.8***		
				(75.70)		
entropy_tag_auto_sn					1,989.2***	
					(560.2)	
Entropy (auto-tag-based, smooth)						78.65***
						(22.15)
Regular savings	272.9***	272.9***	272.9***	274.7***	274.7***	274.7***
	(51.61)	(51.61)	(51.61)	(51.82)	(51.82)	(51.82)
Credit spend	-199.6*	-199.6*	-199.6*	-131.9	-131.9	-131.9
	(110.1)	(110.1)	(110.1)	(109.3)	(109.3)	(109.3)
Month spend	232.8***	232.8***	232.8***	244.1***	244.1***	244.1***
	(16.42)	(16.42)	(16.42)	(17.01)	(17.01)	(17.01)
Urban	426.9	426.9	426.9	416.8	416.8	416.8
	(279.8)	(279.8)	(279.8)	(272.3)	(272.3)	(272.3)
Month income ('000)	-14.94	-14.94	-14.94	-8.932	-8.932	-8.932
	(14.82)	(14.82)	(14.82)	(15.01)	(15.01)	(15.01)
Regular income	-63.08**	-63.08**	-63.08**	-62.21**	-62.21**	-62.21**
	(30.89)	(30.89)	(30.89)	(30.90)	(30.90)	(30.90)
Has income in month	-12.37	-12.37	-12.37	-4.373	-4.373	-4.373
	(88.56)	(88.56)	(88.56)	(87.90)	(87.90)	(87.90)
Loan repayment	-122.7***	-122.7***	-122.7***	-115.9***	-115.9***	-115.9***
	(34.18)	(34.18)	(34.18)	(33.91)	(33.91)	(33.91)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.29141	0.29141	0.29141	0.29173	0.29173	0.29173
Within R^2	0.05498	0.05498	0.05498	0.05541	0.05541	0.05541

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 7: Main results

Dependent Variable:	sa_inflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_merchant	39.13					
	(25.72)					
entropy_merchant_n		394.6				
		(259.4)				
Entropy (merchant-based)			40.06			
			(26.34)			
entropy_merchant_s				902.1**		
				(426.4)		
entropy_merchant_sn					9,098.5**	
					(4,300.6)	
Entropy (merchant-based, smooth)						55.30**
						(26.14)
Regular savings	273.6***	273.6***	273.6***	274.2***	274.2***	274.2***
	(51.68)	(51.68)	(51.68)	(51.96)	(51.96)	(51.96)
Credit spend	-198.1*	-198.1*	-198.1*	-142.1	-142.1	-142.1
	(110.2)	(110.2)	(110.2)	(109.7)	(109.7)	(109.7)
Month spend	234.2***	234.2***	234.2***	243.4***	243.4***	243.4***
	(16.48)	(16.48)	(16.48)	(16.96)	(16.96)	(16.96)
Urban	425.3	425.3	425.3	420.2	420.2	420.2
	(275.7)	(275.7)	(275.7)	(273.4)	(273.4)	(273.4)
Month income ('000)	-14.34	-14.34	-14.34	-9.324	-9.324	-9.324
	(14.80)	(14.80)	(14.80)	(15.12)	(15.12)	(15.12)
Regular income	-62.32**	-62.32**	-62.32**	-61.92**	-61.92**	-61.92**
	(30.87)	(30.87)	(30.87)	(30.99)	(30.99)	(30.99)
Has income in month	-10.71	-10.71	-10.71	-6.934	-6.934	-6.934
	(88.73)	(88.73)	(88.73)	(87.91)	(87.91)	(87.91)
Loan repayment	-118.1***	-118.1***	-118.1***	-115.9***	-115.9***	-115.9***
	(34.15)	(34.15)	(34.15)	(33.92)	(33.92)	(33.92)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.29121	0.29121	0.29121	0.29155	0.29155	0.29155
Within R^2	0.05471	0.05471	0.05471	0.05517	0.05517	0.05517

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 8: Main results

Dependent Variable:			sa_inflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_groc	-9.572					
	(17.94)					
entropy_groc_n		-31.80				
		(59.61)				
entropy_groc_z			-6.503			
			(12.19)			
entropy_groc_s				98.61**		
				(43.19)		
entropy_groc_sn					327.6**	
					(143.5)	
entropy_groc_sz						34.44**
						(15.09)
Regular savings	274.5***	274.5***	274.5***	275.1***	275.1***	275.1***
	(51.83)	(51.83)	(51.83)	(51.85)	(51.85)	(51.85)
Credit spend	-189.2*	-189.2*	-189.2*	-175.2	-175.2	-175.2
	(110.2)	(110.2)	(110.2)	(109.5)	(109.5)	(109.5)
Month spend	235.8***	235.8***	235.8***	237.2***	237.2***	237.2***
	(16.36)	(16.36)	(16.36)	(16.43)	(16.43)	(16.43)
Urban	417.7	417.7	417.7	428.0	428.0	428.0
	(270.7)	(270.7)	(270.7)	(274.6)	(274.6)	(274.6)
Month income ('000)	-13.79	-13.79	-13.79	-12.76	-12.76	-12.76
	(14.82)	(14.82)	(14.82)	(14.91)	(14.91)	(14.91)
Regular income	-62.23**	-62.23**	-62.23**	-61.61**	-61.61**	-61.61**
	(30.85)	(30.85)	(30.85)	(30.94)	(30.94)	(30.94)
Has income in month	-6.875	-6.875	-6.875	-3.608	-3.608	-3.608
	(88.51)	(88.51)	(88.51)	(88.28)	(88.28)	(88.28)
Loan repayment	-115.1***	-115.1***	-115.1***	-114.8***	-114.8***	-114.8***
	(34.10)	(34.10)	(34.10)	(34.03)	(34.03)	(34.03)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.29115	0.29115	0.29115	0.29130	0.29130	0.29130
Within R^2	0.05465	0.05465	0.05465	0.05484	0.05484	0.05484

Clustered (User id) standard-errors in parentheses
Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table 9:   Main results

Dependent Variable:			sa_netflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag	54.32					
	(44.63)					
entropy_tag_n		172.2				
		(141.5)				
Entropy (tag-based)			30.03			
			(24.67)			
entropy_tag_s				128.8**		
				(60.33)		
entropy_tag_sn					408.4**	
					(191.2)	
Entropy (tag-based, smooth)						34.36**
						(16.09)
Regular savings	161.3***	161.3***	161.3***	161.2***	161.2***	161.2***
	(35.23)	(35.23)	(35.23)	(35.24)	(35.24)	(35.24)
Credit spend	-92.56	-92.56	-92.56	-83.60	-83.60	-83.60
	(108.8)	(108.8)	(108.8)	(108.5)	(108.5)	(108.5)
Month spend	128.8***	128.8***	128.8***	129.7***	129.7***	129.7***
	(14.53)	(14.53)	(14.53)	(14.57)	(14.57)	(14.57)
Urban	-250.3	-250.3	-250.3	-250.5	-250.5	-250.5
	(358.6)	(358.6)	(358.6)	(359.3)	(359.3)	(359.3)
Month income ('000)	-126.1***	-126.1***	-126.1***	-125.6***	-125.6***	-125.6***
	(16.86)	(16.86)	(16.86)	(16.85)	(16.85)	(16.85)
Regular income	19.79	19.79	19.79	20.17	20.17	20.17
	(24.59)	(24.59)	(24.59)	(24.56)	(24.56)	(24.56)
Has income in month	174.7*	174.7*	174.7*	176.7*	176.7*	176.7*
	(96.61)	(96.61)	(96.61)	(96.42)	(96.42)	(96.42)
Loan repayment	-77.46***	-77.46***	-77.46***	-77.88***	-77.88***	-77.88***
	(27.25)	(27.25)	(27.25)	(27.20)	(27.20)	(27.20)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.06435	0.06435	0.06435	0.06444	0.06444	0.06444
Within R^2	0.01704	0.01704	0.01704	0.01713	0.01713	0.01713

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 10: Main results

Dependent Variable:	sa_netflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag_auto	53.53*					
	(29.18)					
entropy_tag_auto_n		396.2*				
		(215.9)				
Entropy (auto-tag-based)			47.24*			
			(25.74)			
entropy_tag_auto_s				231.0***		
				(64.05)		
entropy_tag_auto_sn					1,709.6***	
					(474.0)	
Entropy (auto-tag-based, smooth)						67.60***
						(18.74)
Regular savings	161.0***	161.0***	161.0***	162.1***	162.1***	162.1***
	(35.22)	(35.22)	(35.22)	(35.14)	(35.14)	(35.14)
Credit spend	-96.83	-96.83	-96.83	-41.46	-41.46	-41.46
	(108.5)	(108.5)	(108.5)	(107.7)	(107.7)	(107.7)
Month spend	127.6***	127.6***	127.6***	136.4***	136.4***	136.4***
	(14.63)	(14.63)	(14.63)	(15.33)	(15.33)	(15.33)
Urban	-248.3	-248.3	-248.3	-254.2	-254.2	-254.2
	(358.9)	(358.9)	(358.9)	(357.9)	(357.9)	(357.9)
Month income ('000)	-126.4***	-126.4***	-126.4***	-121.6***	-121.6***	-121.6***
	(16.83)	(16.83)	(16.83)	(16.90)	(16.90)	(16.90)
Regular income	19.16	19.16	19.16	19.65	19.65	19.65
	(24.57)	(24.57)	(24.57)	(24.39)	(24.39)	(24.39)
Has income in month	174.3*	174.3*	174.3*	179.6*	179.6*	179.6*
	(96.58)	(96.58)	(96.58)	(95.87)	(95.87)	(95.87)
Loan repayment	-79.83***	-79.83***	-79.83***	-76.32***	-76.32***	-76.32***
	(27.31)	(27.31)	(27.31)	(26.96)	(26.96)	(26.96)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.06440	0.06440	0.06440	0.06476	0.06476	0.06476
Within R^2	0.01709	0.01709	0.01709	0.01747	0.01747	0.01747

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 11: Main results

Dependent Variable:			sa_netflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_merchant	6.463					
	(24.20)					
entropy_merchant_n		65.19				
		(244.1)				
Entropy (merchant-based)			6.618			
			(24.78)			
entropy_merchant_s				1,135.9***		
				(297.8)		
entropy_merchant_sn					11,457.0***	
					(3,003.6)	
Entropy (merchant-based, smooth)						69.64***
						(18.26)
Regular savings	161.6***	161.6***	161.6***	161.6***	161.6***	161.6***
	(35.19)	(35.19)	(35.19)	(35.18)	(35.18)	(35.18)
Credit spend	-93.05	-93.05	-93.05	-30.86	-30.86	-30.86
	(108.6)	(108.6)	(108.6)	(107.4)	(107.4)	(107.4)
Month spend	128.9***	128.9***	128.9***	138.9***	138.9***	138.9***
	(14.73)	(14.73)	(14.73)	(15.57)	(15.57)	(15.57)
Urban	-252.0	-252.0	-252.0	-250.3	-250.3	-250.3
	(357.3)	(357.3)	(357.3)	(359.0)	(359.0)	(359.0)
Month income ('000)	-125.9***	-125.9***	-125.9***	-120.1***	-120.1***	-120.1***
	(16.86)	(16.86)	(16.86)	(16.90)	(16.90)	(16.90)
Regular income	19.60	19.60	19.60	20.04	20.04	20.04
	(24.59)	(24.59)	(24.59)	(24.31)	(24.31)	(24.31)
Has income in month	176.5*	176.5*	176.5*	177.6*	177.6*	177.6*
	(96.59)	(96.59)	(96.59)	(95.63)	(95.63)	(95.63)
Loan repayment	-76.12***	-76.12***	-76.12***	-76.68***	-76.68***	-76.68***
	(27.21)	(27.21)	(27.21)	(26.87)	(26.87)	(26.87)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.06432	0.06432	0.06432	0.06496	0.06496	0.06496
Within R^2	0.01700	0.01700	0.01700	0.01768	0.01768	0.01768

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 12: Main results

Dependent Variable:			sa_netflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_groc	-10.75					
	(17.04)					
entropy_groc_n		-35.69				
		(56.61)				
entropy_groc_z			-7.300			
			(11.58)			
entropy_groc_s				125.9***		
				(38.26)		
entropy_groc_sn					418.2***	
					(127.1)	
entropy_groc_sz						43.97***
						(13.36)
Regular savings	161.9***	161.9***	161.9***	162.7***	162.7***	162.7***
	(35.22)	(35.22)	(35.22)	(35.20)	(35.20)	(35.20)
Credit spend	-90.37	-90.37	-90.37	-72.24	-72.24	-72.24
	(108.8)	(108.8)	(108.8)	(108.8)	(108.8)	(108.8)
Month spend	129.3***	129.3***	129.3***	131.1***	131.1***	131.1***
	(14.55)	(14.55)	(14.55)	(14.62)	(14.62)	(14.62)
Urban	-253.5	-253.5	-253.5	-240.4	-240.4	-240.4
	(356.4)	(356.4)	(356.4)	(359.1)	(359.1)	(359.1)
Month income ('000)	-125.8***	-125.8***	-125.8***	-124.4***	-124.4***	-124.4***
	(16.87)	(16.87)	(16.87)	(16.85)	(16.85)	(16.85)
Regular income	19.64	19.64	19.64	20.43	20.43	20.43
	(24.59)	(24.59)	(24.59)	(24.53)	(24.53)	(24.53)
Has income in month	177.6*	177.6*	177.6*	181.8*	181.8*	181.8*
	(96.47)	(96.47)	(96.47)	(96.43)	(96.43)	(96.43)
Loan repayment	-75.60***	-75.60***	-75.60***	-75.25***	-75.25***	-75.25***
	(27.10)	(27.10)	(27.10)	(27.03)	(27.03)	(27.03)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.06432	0.06432	0.06432	0.06457	0.06457	0.06457
Within R^2	0.01701	0.01701	0.01701	0.01727	0.01727	0.01727

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 13: Main results

Dependent Variable:	sa_outflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag	8.360					
	(47.32)					
entropy_tag_n		26.50				
		(150.0)				
Entropy (tag-based)			4.621			
			(26.16)			
entropy_tag_s				-14.43		
				(62.44)		
entropy_tag_sn					-45.76	
					(197.9)	
Entropy (tag-based, smooth)						-3.849
						(16.65)
Regular savings	112.5**	112.5**	112.5**	112.6**	112.6**	112.6**
	(50.58)	(50.58)	(50.58)	(50.60)	(50.60)	(50.60)
Credit spend	-98.83	-98.83	-98.83	-99.63	-99.63	-99.63
	(108.7)	(108.7)	(108.7)	(108.6)	(108.6)	(108.6)
Month spend	106.5***	106.5***	106.5***	106.5***	106.5***	106.5***
	(11.99)	(11.99)	(11.99)	(12.03)	(12.03)	(12.03)
Urban	671.6***	671.6***	671.6***	670.9***	670.9***	670.9***
	(218.1)	(218.1)	(218.1)	(217.9)	(217.9)	(217.9)
Month income ('000)	111.9***	111.9***	111.9***	112.0***	112.0***	112.0***
	(21.79)	(21.79)	(21.79)	(21.79)	(21.79)	(21.79)
Regular income	-81.84**	-81.84**	-81.84**	-81.93**	-81.93**	-81.93**
	(32.40)	(32.40)	(32.40)	(32.40)	(32.40)	(32.40)
Has income in month	-184.8**	-184.8**	-184.8**	-184.4**	-184.4**	-184.4**
	(90.65)	(90.65)	(90.65)	(90.69)	(90.69)	(90.69)
Loan repayment	-39.75	-39.75	-39.75	-39.21	-39.21	-39.21
	(31.16)	(31.16)	(31.16)	(31.12)	(31.12)	(31.12)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.23999	0.23999	0.23999	0.23999	0.23999	0.23999
Within R^2	0.01892	0.01892	0.01892	0.01892	0.01892	0.01892

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

16

Table 14: Main results

Dependent Variable:	sa_outflows					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_tag_auto	43.32					
	(34.84)					
entropy_tag_auto_n		320.6				
		(257.9)				
Entropy (auto-tag-based)			38.23			
			(30.75)			
entropy_tag_auto_s				37.78		
				(84.25)		
entropy_tag_auto_sn					279.6	
					(623.5)	
Entropy (auto-tag-based, smooth)						11.06
						(24.65)
Regular savings	112.0**	112.0**	112.0**	112.6**	112.6**	112.6**
	(50.49)	(50.49)	(50.49)	(50.63)	(50.63)	(50.63)
Credit spend	-102.8	-102.8	-102.8	-90.48	-90.48	-90.48
	(108.8)	(108.8)	(108.8)	(106.9)	(106.9)	(106.9)
Month spend	105.3***	105.3***	105.3***	107.7***	107.7***	107.7***
	(12.04)	(12.04)	(12.04)	(11.99)	(11.99)	(11.99)
Urban	675.2***	675.2***	675.2***	671.0***	671.0***	671.0***
	(220.6)	(220.6)	(220.6)	(217.7)	(217.7)	(217.7)
Month income ('000)	111.5***	111.5***	111.5***	112.7***	112.7***	112.7***
	(21.78)	(21.78)	(21.78)	(22.33)	(22.33)	(22.33)
Regular income	-82.24**	-82.24**	-82.24**	-81.86**	-81.86**	-81.86**
	(32.45)	(32.45)	(32.45)	(32.45)	(32.45)	(32.45)
Has income in month	-186.7**	-186.7**	-186.7**	-184.0**	-184.0**	-184.0**
	(90.84)	(90.84)	(90.84)	(90.54)	(90.54)	(90.54)
Loan repayment	-42.87	-42.87	-42.87	-39.58	-39.58	-39.58
	(31.04)	(31.04)	(31.04)	(31.16)	(31.16)	(31.16)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.24004	0.24004	0.24004	0.24000	0.24000	0.24000
Within R^2	0.01899	0.01899	0.01899	0.01894	0.01894	0.01894

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 15: Main results

Dependent Variable:			sa_outflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_merchant	32.66					
	(27.07)					
entropy_merchant_n		329.5				
		(273.0)				
Entropy (merchant-based)			33.45			
			(27.72)			
entropy_merchant_s				-233.8		
				(472.4)		
entropy_merchant_sn					-2,358.5	
					(4,764.6)	
Entropy (merchant-based, smooth)						-14.34
						(28.96)
Regular savings	112.0**	112.0**	112.0**	112.6**	112.6**	112.6**
	(50.51)	(50.51)	(50.51)	(50.59)	(50.59)	(50.59)
Credit spend	-105.1	-105.1	-105.1	-111.3	-111.3	-111.3
	(108.7)	(108.7)	(108.7)	(107.1)	(107.1)	(107.1)
Month spend	105.3***	105.3***	105.3***	104.5***	104.5***	104.5***
	(12.11)	(12.11)	(12.11)	(11.61)	(11.61)	(11.61)
Urban	677.3***	677.3***	677.3***	670.6***	670.6***	670.6***
	(220.8)	(220.8)	(220.8)	(217.7)	(217.7)	(217.7)
Month income ('000)	111.5***	111.5***	111.5***	110.8***	110.8***	110.8***
	(21.79)	(21.79)	(21.79)	(22.54)	(22.54)	(22.54)
Regular income	-81.92**	-81.92**	-81.92**	-81.96**	-81.96**	-81.96**
	(32.44)	(32.44)	(32.44)	(32.40)	(32.40)	(32.40)
Has income in month	-187.3**	-187.3**	-187.3**	-184.5**	-184.5**	-184.5**
	(90.82)	(90.82)	(90.82)	(90.60)	(90.60)	(90.60)
Loan repayment	-42.01	-42.01	-42.01	-39.24	-39.24	-39.24
	(31.17)	(31.17)	(31.17)	(31.18)	(31.18)	(31.18)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.24003	0.24003	0.24003	0.24002	0.24002	0.24002
Within R^2	0.01897	0.01897	0.01897	0.01895	0.01895	0.01895

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

18

Table 16: Main results

Dependent Variable:			sa_outflows			
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
entropy_groc	1.173					
	(17.29)					
entropy_groc_n		3.896				
		(57.45)				
entropy_groc_z			0.7968			
			(11.75)			
entropy_groc_s				-27.29		
				(43.01)		
entropy_groc_sn					-90.65	
					(142.9)	
entropy_groc_sz						-9.532
						(15.02)
Regular savings	112.6**	112.6**	112.6**	112.4**	112.4**	112.4**
	(50.65)	(50.65)	(50.65)	(50.55)	(50.55)	(50.55)
Credit spend	-98.87	-98.87	-98.87	-103.0	-103.0	-103.0
	(108.8)	(108.8)	(108.8)	(108.1)	(108.1)	(108.1)
Month spend	106.5***	106.5***	106.5***	106.1***	106.1***	106.1***
	(12.04)	(12.04)	(12.04)	(11.93)	(11.93)	(11.93)
Urban	671.2***	671.2***	671.2***	668.4***	668.4***	668.4***
	(217.7)	(217.7)	(217.7)	(217.4)	(217.4)	(217.4)
Month income ('000)	112.0***	112.0***	112.0***	111.7***	111.7***	111.7***
	(21.79)	(21.79)	(21.79)	(21.87)	(21.87)	(21.87)
Regular income	-81.87**	-81.87**	-81.87**	-82.05**	-82.05**	-82.05**
	(32.41)	(32.41)	(32.41)	(32.47)	(32.47)	(32.47)
Has income in month	-184.5**	-184.5**	-184.5**	-185.4**	-185.4**	-185.4**
	(90.73)	(90.73)	(90.73)	(90.49)	(90.49)	(90.49)
Loan repayment	-39.46	-39.46	-39.46	-39.54	-39.54	-39.54
	(31.19)	(31.19)	(31.19)	(31.21)	(31.21)	(31.21)
Fixed-effects						
User id	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	49,636	49,636	49,636	49,636	49,636	49,636
R^2	0.23999	0.23999	0.23999	0.24000	0.24000	0.24000
Within R^2	0.01892	0.01892	0.01892	0.01894	0.01894	0.01894

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

- Smoothed entropy is high when the number of unique categories is low (and zero categories high), and it is negatively related to saving. Hence: user-months with spends across few categories are less likely to save.

Unsmoothed entropy is high when spends across categories are equal and the total number of spends is high.

Table 17: FE specifications

Dependent Variable:	has_sa_inflows		
Model:	(1)	(2)	(3)
Variables			
Entropy (tag-based, smooth)	-0.0509***	-0.0500***	-0.0380***
	(0.0033)	(0.0067)	(0.0067)
Entropy (tag-based)	0.0960***	0.0881***	0.0656***
	(0.0052)	(0.0104)	(0.0106)
Regular savings	0.6200***	0.2150***	0.2097***
	(0.0046)	(0.0088)	(0.0088)
Credit spend	-0.1115***	-0.0384*	-0.0353*
	(0.0082)	(0.0199)	(0.0198)
Month spend	0.0078***	0.0202***	0.0194***
	(0.0007)	(0.0013)	(0.0013)
Urban	0.0129***	0.0392	0.0368
	(0.0044)	(0.1268)	(0.1209)
Month income ('000)	-0.0081***	0.0031	0.0022
	(0.0009)	(0.0019)	(0.0019)
Regular income	-0.1203***	-0.0109*	-0.0258***
	(0.0034)	(0.0064)	(0.0075)
Has income in month	0.1228***	0.0180	0.0196
	(0.0133)	(0.0167)	(0.0166)
Loan repayment	-0.0022	0.0049	0.0025
	(0.0033)	(0.0085)	(0.0084)
(Intercept)	0.2450***		
	(0.0136)		
Fixed-effects			
User id		Yes	Yes
Calendar month			Yes
Fit statistics			
Observations	82,589	82,589	82,589
R^2	0.19349	0.41341	0.41652
Within R^2		0.03498	0.02849

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

Table 18: FE specifications

Dependent Variable: Model:	(1)	(2)	(3)	(4)	has_sa_inflows (5)	(6)	(7)	(8)	(9)
Variables									
Entropy (tag-based, smooth)	-0.0612***	-0.0587***	-0.0651***	-0.0369***	-0.0612***	-0.0593***	-0.0608***	-0.0607***	-0.0607***
	(0.0069)	(0.0066)	(0.0071)	(0.0068)	(0.0069)	(0.0068)	(0.0069)	(0.0069)	(0.0068)
Entropy (tag-based)	0.1037***	0.0989***	0.1090***	0.0658***	0.1037***	0.1003***	0.1030***	0.1027***	0.1023***
	(0.0109)	(0.0104)	(0.0113)	(0.0108)	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0109)
Regular savings		0.2020***							
		(0.0090)							
Credit spend			-0.0391*						
			(0.0211)						
Month spend				0.0196***					
				(0.0013)					
Urban					0.0007				
					(0.1404)				
Month income ('000)						0.0057***			
						(0.0020)			
Regular income							0.0101		
							(0.0077)		
Has income in month								0.0229	
								(0.0172)	
Loan repayment									0.0147
									(0.0090)
Fixed-effects									
User id	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Calendar month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	82,589	82,589	82,589	82,589	82,589	82,589	82,589	82,589	82,589
R²	0.40153	0.41216	0.40162	0.40548	0.40153	0.40170	0.40158	0.40156	0.40162
Within R²	0.00354	0.02123	0.00368	0.01012	0.00354	0.00382	0.00362	0.00359	0.00368

Clustered (User id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 1: Entropy and unique tags
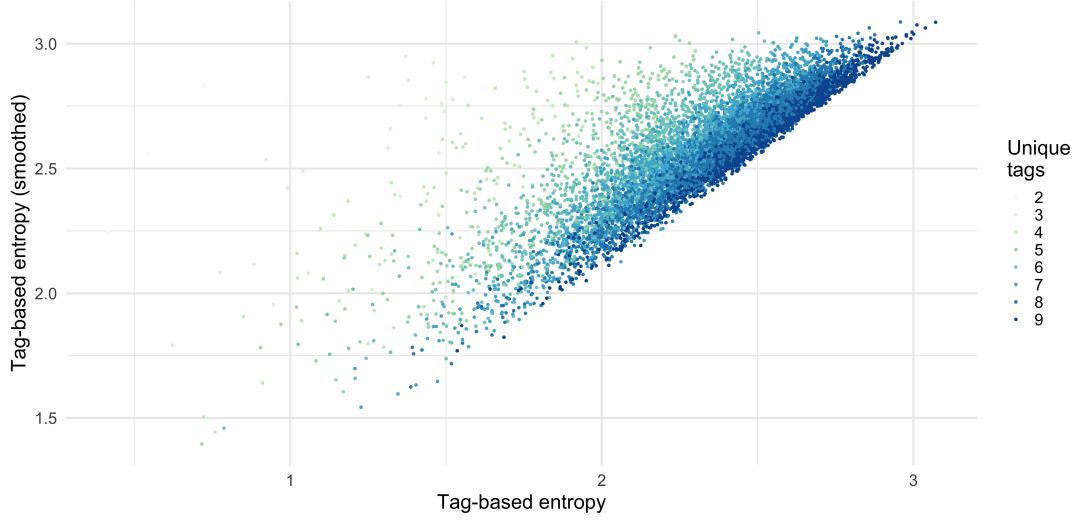
3 Discussion

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]

---

[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).

4.2   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 19 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 19: Sample selection

	Users	User-months	Txns	Txns (m£)
Raw sample	26,513	739,047	64,464,494	11,924.50
Annual income of at least £10k	8,154	199,274	20,381,666	3,644.02
Income in 2/3 of all observed months	8,032	197,422	20,213,950	3,609.57
At least one savings account	4,751	127,958	13,776,356	2,660.75
At least 6 months of data	4,283	126,703	13,668,274	2,645.28
Monthly debits of at least £200	3,555	103,604	11,670,234	2,218.18
Five or more monthly current account txns	3,312	96,588	10,771,752	2,006.45
Spends in two distinct categories each month	3,276	95,350	10,668,163	1,982.41
Complete demographic information	2,730	82,589	9,266,079	1,709.51
Final sample	2,730	82,589	9,266,079	1,709.51

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

---

[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.

- 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.
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.

---

[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.

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

Figure 2: Transactions distributions
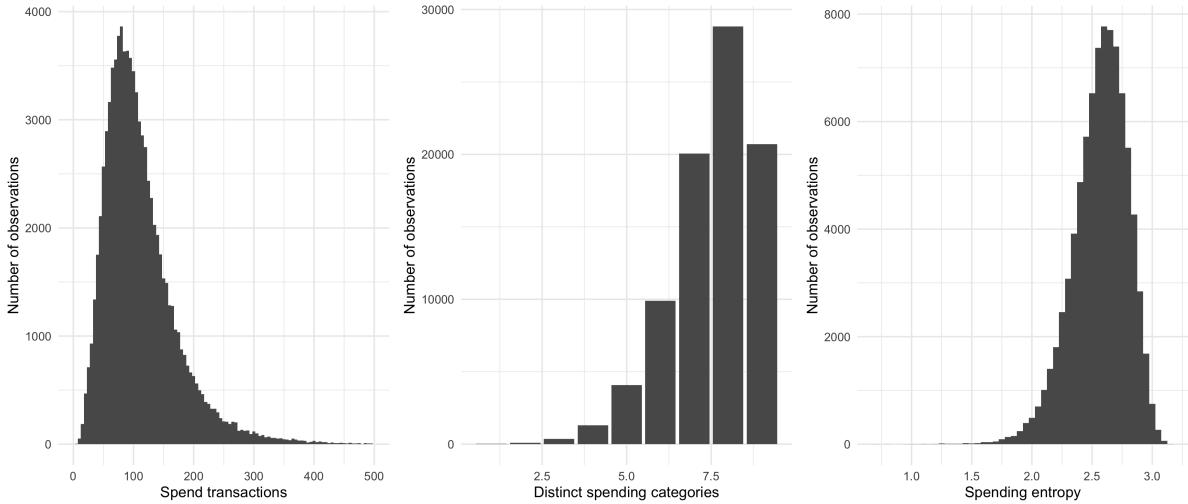
Notes: From top-left to bottom-right: distribution of spending transactions per user-month, breakdown of spending transactions into spending categories; breakdown of number of spending categories spent on in user-month; distribution of user-month entropy scores.

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 ...

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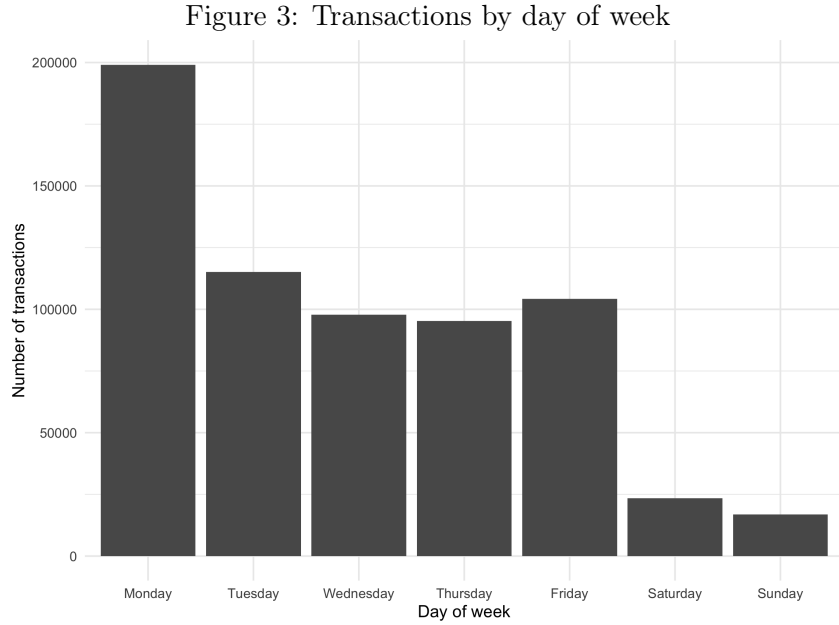
Notes: Number of transactions by day of week, based on a 1/1000 sample of the full data. Shows that banks process most weekend transactions on Mondays.

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   Summary statistics

Table 20 provides summary statistics.

Figure 4

Table 20: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
user_id	49,636	352,410.200	170,148.300	507	244,047	487,367	589,517
ym	49,636	574.942	18.927	505	565	589	606
txns_count	49,636	120.758	62.054	7	79	146	756
txns_volume	49,636	18,159.330	63,659.210	256.890	5,886.788	16,937.940	10,034,724.000
txn_count_sa	49,636	4.642	9.404	0	0	6	389
txn_count_ca	49,636	103.516	55.317	5	67	127	738
nunique_tag	49,636	7.774	1.070	2	7	9	9
nunique_tag_auto	49,636	22.230	6.701	3	17	27	50
nunique_merchant	49,636	25.398	9.110	2	19	31	73
prop_credit	49,636	0.080	0.176	0.000	0.000	0.041	0.999
month_income	49,636	2.695	1.745	0.000	1.568	3.342	9.249
year_income	49,636	32.109	17.366	10.000	19.507	39.733	86.380
has_regular_income	49,636	0.630	0.483	0	0	1	1
has_month_income	49,636	0.988	0.107	0	1	1	1
age	49,636	36.574	10.588	17	29	42	89
is_female	49,636	0.416	0.493	0	0	1	1
sa_inflows	49,636	642.931	1,756.068	0	0	450	11,430
sa_outflows	49,636	636.611	1,755.491	0	0	470	11,500
sa_netflows	49,636	6.320	1,725.718	−11,500	−80	0	11,430
has_sa_inflows	49,636	0.443	0.497	0	0	1	1
has_reg_sa_inflows	49,636	0.155	0.362	0	0	0	1
has_benefits	49,636	0.067	0.250	0	0	0	1
has_pension	49,636	0.022	0.146	0	0	0	1
has_rent_payment	49,636	0.122	0.327	0	0	0	1
has_mortgage_payment	49,636	0.221	0.415	0	0	0	1
has_loan_funds	49,636	0.028	0.166	0	0	0	1
has_loan_repmt	49,636	0.370	0.483	0	0	1	1
is_urban	49,636	0.848	0.359	0	1	1	1
spend_services	49,636	0.457	0.457	0.000	0.179	0.561	2.835
spend_other_spend	49,636	0.282	0.365	0.000	0.062	0.354	2.246
spend_household	49,636	1.039	1.022	0.0004	0.365	1.354	6.104
spend_travel	49,636	0.136	0.301	0	0	0.1	2
spend_communication	49,636	0.071	0.059	0.000	0.030	0.093	0.300
spend_finance	49,636	0.312	0.493	0.000	0.034	0.374	3.127
spend_motor	49,636	0.094	0.103	0.000	0.008	0.138	0.468
spend_retail	49,636	0.167	0.212	0	0.03	0.2	1
spend_hobbies	49,636	0.026	0.052	0	0	0.03	0
month_spend	49,636	2.725	2.346	0.202	1.297	3.304	14.954
has_od_fees	49,636	0.242	0.429	0	0	0	1
entropy_tag	49,636	2.313	0.282	0.715	2.147	2.514	3.071
entropy_tag_n	49,636	0.730	0.089	0.226	0.677	0.793	0.969
entropy_tag_z	49,636	0.380	0.510	−2.510	0.080	0.743	1.752
entropy_tag_s	49,636	2.541	0.215	1.365	2.413	2.694	3.118
entropy_tag_sn	49,636	0.802	0.068	0.431	0.761	0.850	0.983
entropy_tag_sz	49,636	−0.205	0.805	−4.618	−0.688	0.367	1.955
entropy_tag_auto	49,636	3.571	0.496	0.979	3.263	3.928	5.038
entropy_tag_auto_n	49,636	0.483	0.067	0.132	0.441	0.531	0.681
entropy_tag_auto_z	49,636	0.466	0.562	−2.472	0.117	0.871	2.128
entropy_tag_auto_s	49,636	6.875	0.281	4.245	6.735	7.078	7.380
entropy_tag_auto_sn	49,636	0.929	0.038	0.574	0.910	0.956	0.997
entropy_tag_auto_sz	49,636	−0.490	0.960	−9.477	−0.971	0.201	1.234
entropy_merchant	49,636	3.986	0.586	0.353	3.632	4.400	5.696
entropy_merchant_n	49,636	0.395	0.058	0.035	0.360	0.436	0.565
entropy_merchant_z	49,636	0.530	0.572	−3.018	0.184	0.934	2.200
entropy_merchant_s	49,636	10.004	0.066	9.026	9.985	10.046	10.085
entropy_merchant_sn	49,636	0.992	0.007	0.895	0.990	0.996	1.000
entropy_merchant_sz	49,636	−0.396	1.076	−16.357	−0.715	0.286	0.912
entropy_groc	49,636	1.358	0.614	0.000	0.971	1.823	3.003
entropy_groc_n	49,636	0.409	0.185	0.000	0.292	0.549	0.904
entropy_groc_z	49,636	0.333	0.904	−1.665	−0.236	1.018	2.754
entropy_groc_s	49,636	2.731	0.339	0.603	2.553	2.977	3.278
entropy_groc_sn	49,636	0.822	0.102	0.182	0.768	0.896	0.987
entropy_groc_sz	49,636	−0.305	0.971	−6.396	−0.815	0.400	1.260

Figure 4: Entropy distributions
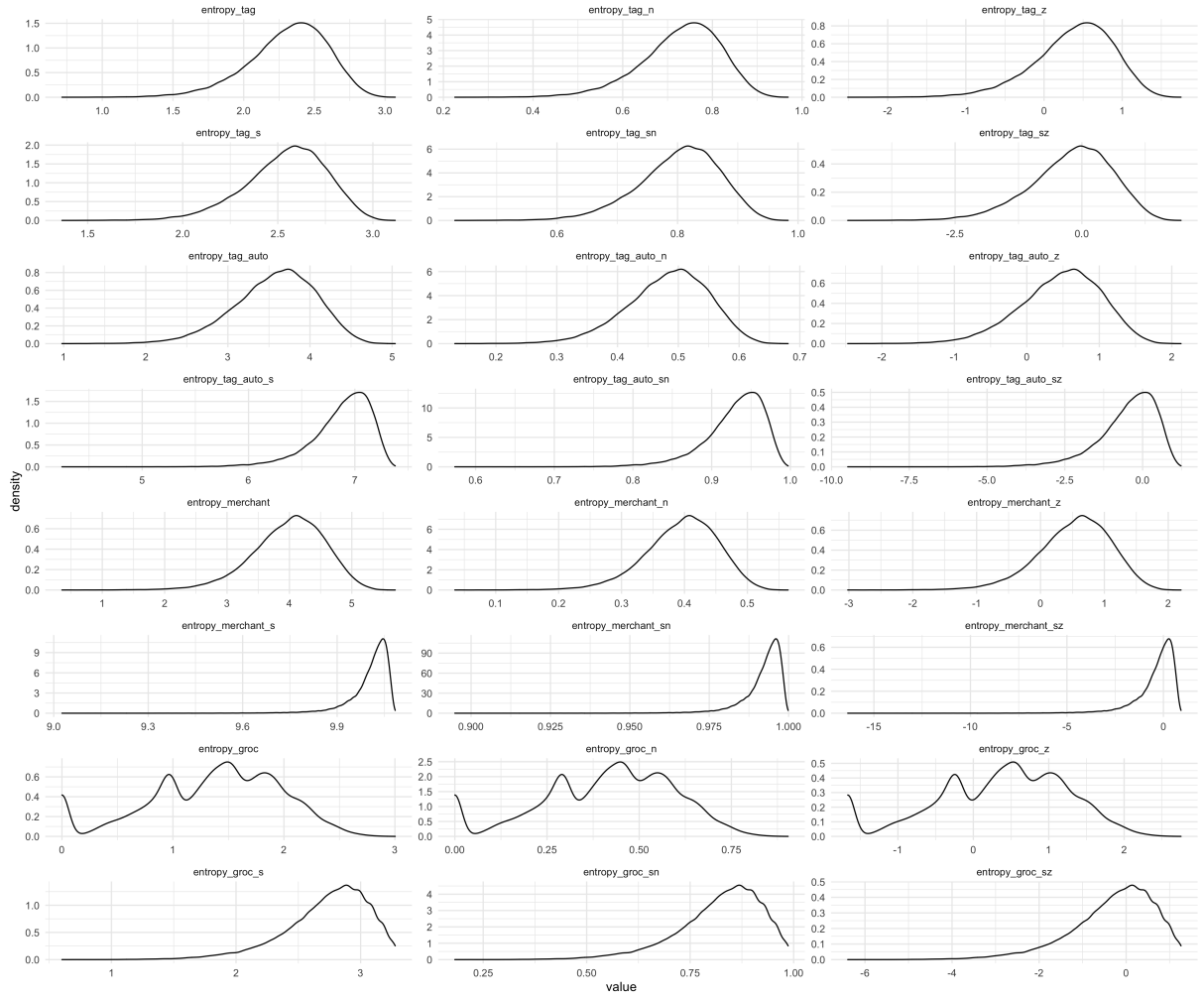
Notes:

4.6 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.

28

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 Variable construction

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 –

Individual and household characteristics

- Gender

- Age

- Urban

- Region

- Month income, winsorised at top 1 percent-level.

- Year income, winsorised at the 1 percent level.

- Regular income, dummy for 10 out of last 12 months

---

[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.

- 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 –

A.3   Variabel description

Figure 5: Demographic characteristics of Money Dashboard users
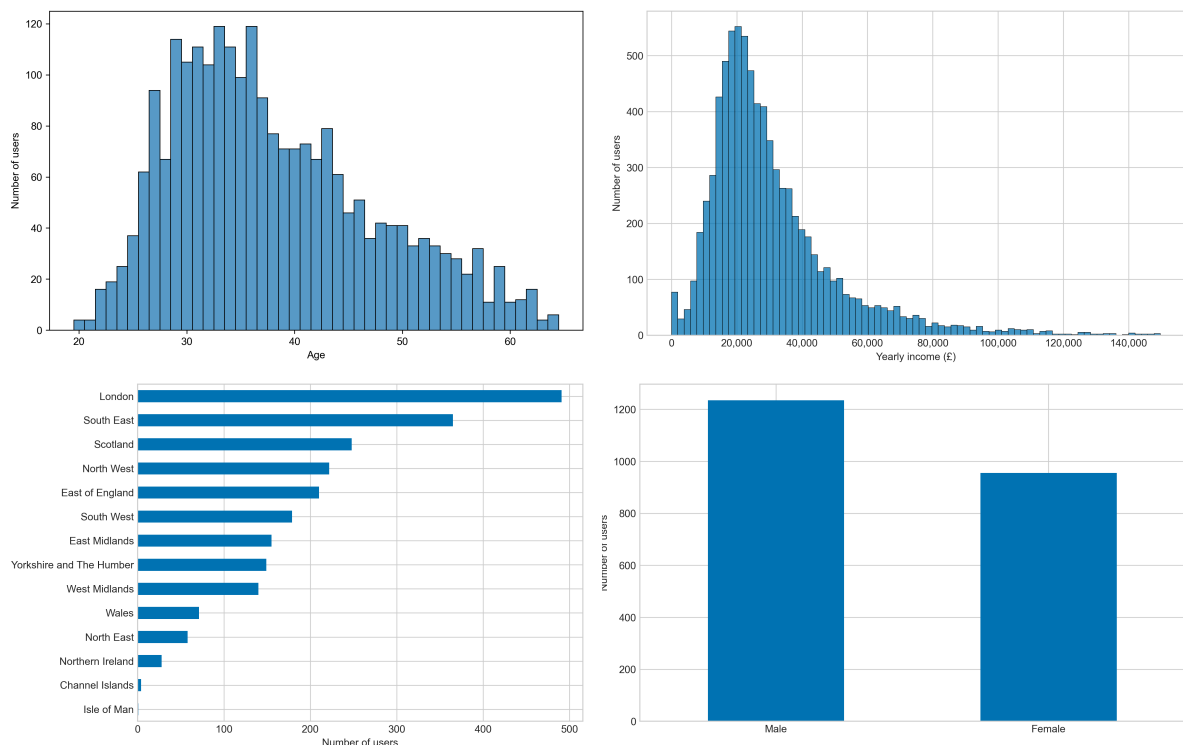
Exploration of control variables here (e.g. like jpmorgan2019weathering for income stability)

A.4   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 21 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.

Table 21: Effect of entropy on overdraft fees

Dependent Variable:		has_od_fees		
Model:	(1)	(2)	(3)	(4)
Variables				
Entropy (tag-based, smooth)	0.0040**		0.0052*	
	(0.0018)		(0.0027)	
Entropy (tag-based)		0.0745***		0.0318***
		(0.0029)		(0.0044)
Spend communication	0.6974***	0.5952***	0.1294***	0.1025***
	(0.0256)	(0.0258)	(0.0385)	(0.0384)
Spend finance	0.0256***	0.0192***	0.0124***	0.0110***
	(0.0028)	(0.0028)	(0.0037)	(0.0037)
Spend hobbies	0.0040	-0.1075***	0.0411	0.0106
	(0.0283)	(0.0284)	(0.0263)	(0.0265)
Spend household	-0.0011	0.0018	0.0027	0.0033
	(0.0016)	(0.0016)	(0.0021)	(0.0021)
Spend other	0.0188***	0.0072*	-0.0053	-0.0091**
	(0.0040)	(0.0040)	(0.0040)	(0.0040)
Spend motor	0.1373***	0.0441***	0.0326	0.0068
	(0.0154)	(0.0157)	(0.0208)	(0.0208)
Spend retail	0.0109	-0.0124*	0.0018	-0.0063
	(0.0073)	(0.0074)	(0.0068)	(0.0070)
Spend services	-0.0073**	0.0085**	-0.0054	-0.0026
	(0.0036)	(0.0035)	(0.0039)	(0.0038)
Spend travel	-0.0329***	-0.0449***	-0.0041	-0.0078**
	(0.0049)	(0.0049)	(0.0036)	(0.0036)
Female	0.0429***	0.0367***	-0.3720	-0.3891
	(0.0030)	(0.0030)	(47,268.6)	(47,267.8)
Age	-0.0026***	-0.0026***		
	(0.0001)	(0.0001)		
Year income	-0.0015***	-0.0015***	0.0005*	0.0005
	(9.26×10^{-5})	(9.22×10^{-5})	(0.0003)	(0.0003)
(Intercept)	0.2793***	0.2746***		
	(0.0057)	(0.0056)		
Fixed-effects				
User id			Yes	Yes
Calendar month			Yes	Yes
Fit statistics				
Observations	82,589	82,589	82,589	82,589
R^2	0.02522	0.03275	0.63492	0.63561
Within R^2			0.00162	0.00352

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