Entropy*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

February 11, 2022

Contents

1 Introduction

We use the following nomenclature throughout:

user individual

tag spending category

Question:

- Do people forget to save when their lives are chaotic?

Definitions:

- We define a savings transaction as an inflow into any of a person's savings accounts.
- We capture the degree of chaos in a person's life using entropy.

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

Motivation:

- Understanding what determines people's savings behviour is important but, to our knowledge, completely understudied.
- Entropy is positively correlated with negative financial outcomes Muggleton et al. (2020)

2 Data

2.1 Dataset description

Data is provided by Money Dashboard (MDB), a UK-based financial management app that allows its users to add accounts from all their banks to obtain an integrated view of their finances. Our dataset contains information on more than 500 million transactions made between 2012 and June 2020 by more than 250,000 users. For each transaction, we can see the amount, date, and description of the transaction, as well as transaction *tags*, classifications added by MDB that indicate the type of the transaction (e.g. 'groceries', 'insurance'). We also have basic information on each user (e.g. year of birth, postcode sector) as well information about each bank account (e.g. type of account, date added).

The main advantages of the data for the study of consumer financial behaviour are its high (transaction-level) 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, MAS 2014). 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, one could argue, for a higher degree of financial sophistication. However, while non-representativeness could partially be addressed by re-weighting the sample, as was done in Bourquin et al. (2020), it is not of much consequence for our purpose here, since our ability to infer behaviour traits from transaction data is not dependent on having a representative sample of people.

Further limitations:

- To the extent that users link shared accounts, they might be more appropriately thought of as households rather than individual usres (Bourquin et al. 2020). We assume that in the majority of cases, shared partner accounts are used for shared household expenses rather than personal expenses, and that salary payments are paid into personal accounts. To the extent that this is true, identified salaries are to a single individual, and expenses made by a partner with a shared account would mainly be for household items that an individual would have also purchased if they lived on their own (albeit in smaller quantities), but not for additional spending categories, which would impact our entropy spending tag based entropy measure.

- 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.

2.2 Preprocessing and sample selection

The MDB data is noisy. We perform a number of preprocessign steps to deal with that.

- 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.

Table 1: Sample selection

	Users	Accounts	Transactions	Value (£M)
Raw sample	26,513	130,058	64,359,662	11,901.7
Annual income of at least £10k	8,156	37,264	20,351,895	3,638.2
Income in 2/3 of all observed months	8,033	36,809	20,174,721	3,602.1
At least one savings account	4,752	28,227	13,748,247	2,655.0
At least 6 months of data	4,284	26,753	13,640,172	2,639.5
Monthly debits of at least £200	3,556	21,907	11,645,522	2,213.3
Five or more current account txns per month	3,312	20,247	10,745,968	2,001.4
Complete demographic information	2,777	17,117	9,371,063	1,730.4
Final sample	2,777	17,117	9,371,063	1,730.4

2.3 Summary statistics

The four panels in Figure 1 provide an overview of demographic characteristics of our sample. It makes clear that Money Dashboard users are not a representative sample of the UK population: they are predominantly males in their thirties who live in London or the South East and are relatively well off (the income distribution is shifted to the right relative to the UK as a whole).[1]

[1] To calculate incomes, we broadly follow Hacioglu et al. (2020) in defining total income as the sum of earnings, pension income, benefits, and other income.

Figure 1: Demographic characteristics of Money Dashboard users

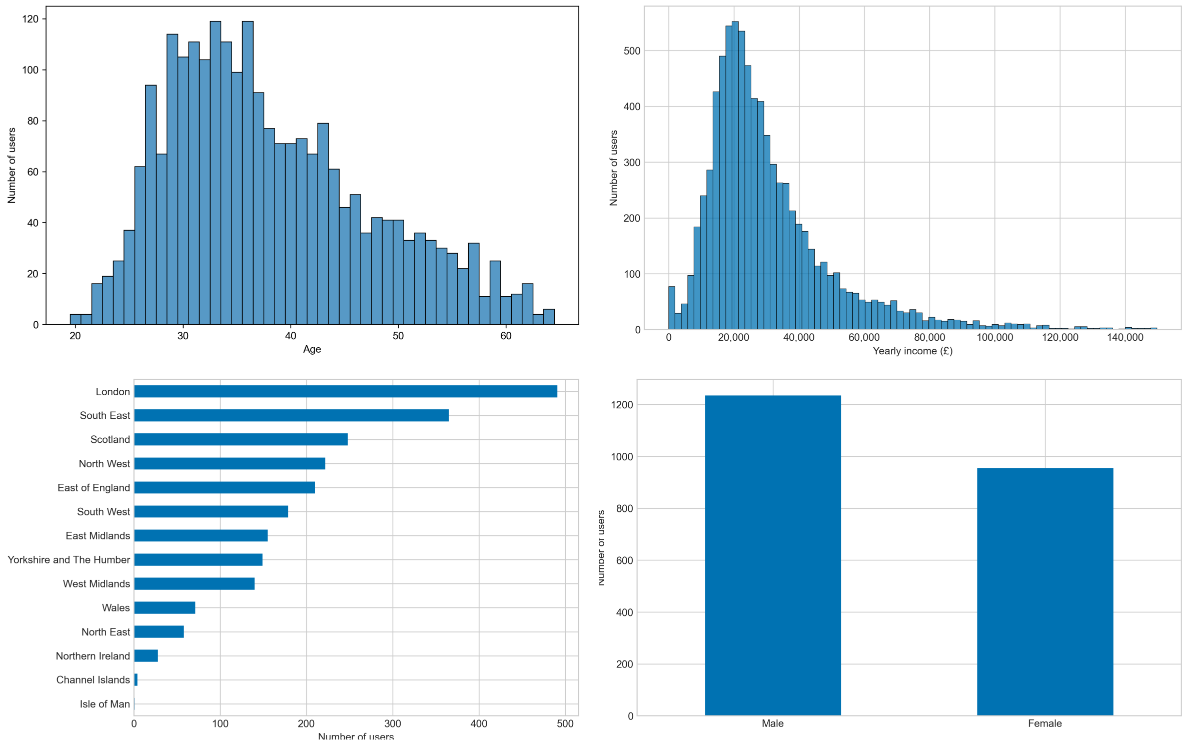
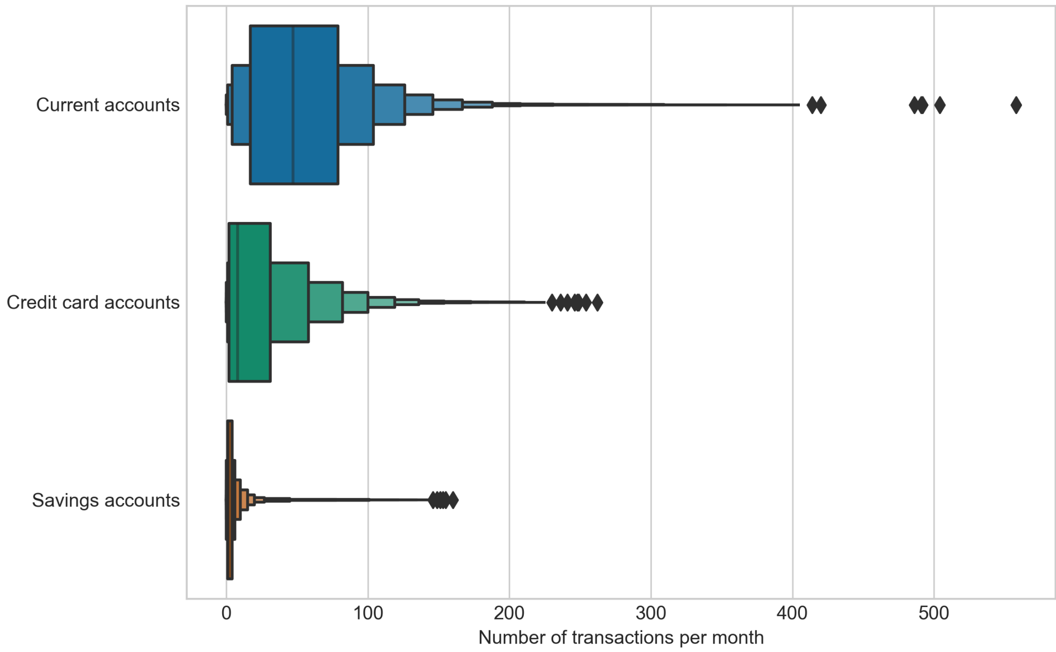


Figure 2: Monthly transactions by account type



Notes: The two innermost boxes in the letter-value plots are identical to those in a boxplot, with the center line corresponding to the median and the left and right edges to the first and third quartiles, respectively – or half of the remaining data on either side of the median. Additional boxes on either side extend that principle by corresponding to half of the remaining data on that side. For instance, the second box to the right of the median in the current accounts plot indicates that half of all account-month observations to the right of the third quartile have fewer than about 105 transactions. Boxes of the same height correspond to the same level, individually drawn observations are outliers.

Table 2: Summary statistics

	count	mean	std	min	max	25%	
obs	163915	98.0689	51.2574	11	656	63	
balance_ca	159116	1130.2	4834.21	-12257.1	34628.6	-884.008	81
balance_sa	64935	2625.16	5496.73	-1644.62	42168.6	2.21001	
sa_inflows	60583	780.844	1507.55	0	13800	60	
sa_outflows	60583	749.975	1452.1	0	12075.5	0	
sa_net_inflows	61195	75.7461	3330.15	-69750	120000	-180	
sa_scaled_inflows	59977	0.335841	0.548266	0	4.13782	0.0319006	0.1
sa_scaled_outflows	59977	0.332819	0.57624	0	4.03946	0	0.1
sa_scaled_net_inflows	59971	0.00713739	0.614697	-4.03481	4.18434	-0.0845304	0.008
total_monthly_spend	156510	7.28142	0.728324	4.95442	9.16108	6.81671	7.
tag_spend_household	160635	0.349356	0.248708	-0.840723	1.61585	0.174674	0.
tag_spend_hobbies	160635	0.0116385	0.0231921	-0.00630346	0.184913	0	
tag_spend_retail	160635	0.0598939	0.103136	-0.620232	0.694001	0.00454646	0.03
tag_spend_services	160635	0.199177	0.174157	-0.589876	1.08971	0.0882128	0.
tag_spend_other_spend	160635	0.117678	0.196987	-1.08001	1.50045	0.0189057	0.07
tag_spend_finance	160635	0.110878	0.155252	-0.279174	0.871723	0.00624401	0.04
tag_spend_travel	160635	0.0552881	0.0956478	-0.0688958	0.624166	0	0.01
tag_spend_communication	160635	0.0401485	0.0449402	-0.0878466	0.320432	0.0115177	0.02
tag_spend_motor	160635	0.0388507	0.0528382	-0.0443517	0.330336	0	0.01
entropy_sptac	160635	2.56947	0.215097	1.90212	2.99952	2.43239	2.
log_income	163915	10.0934	0.580853	8.51758	12.1774	9.72488	1
user_female	154619	0.412259	0.492243	0	1	0	
age	149907	35.2528	10.9868	15	134	27	

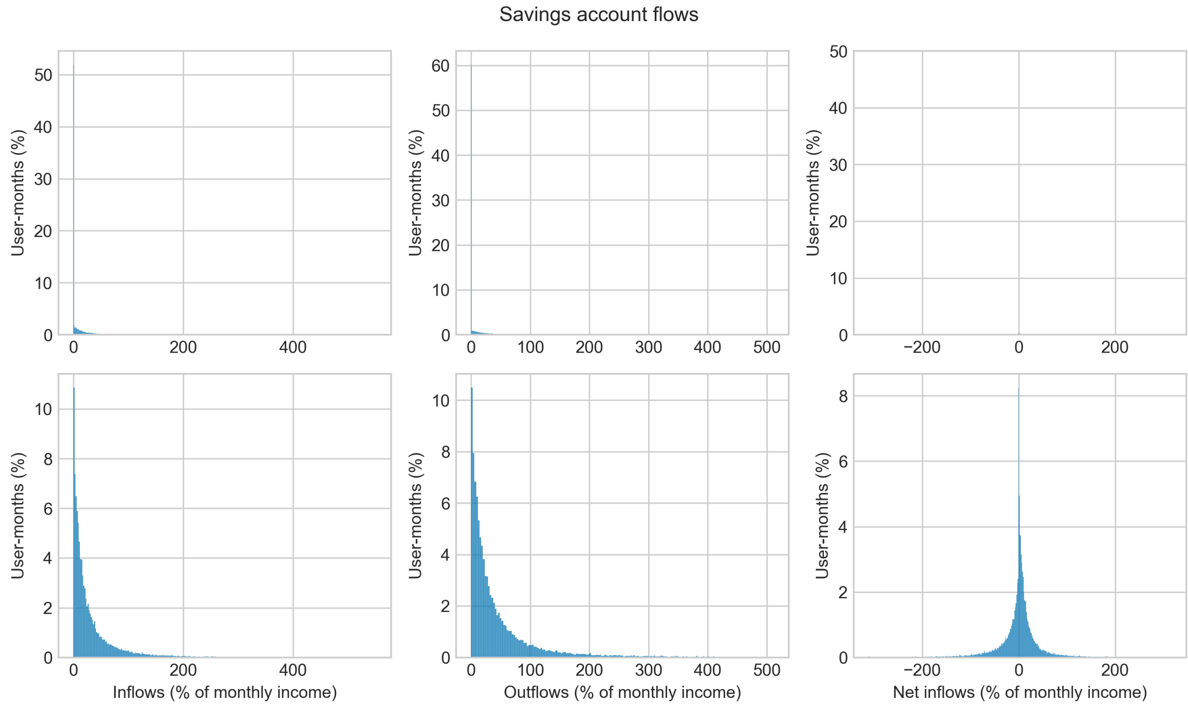
2.4 Dependent variable

Our interest is in short and medium-term savings, which makes savings accounts the natural vehicle to focus on, and our hypothesis is that chaos makes people forget to save, making an indicator of whether or not an individual made a savings transaction in a given period is natural. Hence, our dependent variable is a binary indicator for whether or not an individual made a payment into any of their savings accounts in a perticular period, which we set to a month.

We focus on monthly periods because it is the period into which many people partition their financial lives (mainly because many receive their salary once per month).

We calculate savings account inflows as the sum of all credits that are not identified as interest payments into a user's savings accounts. We would expect that it is particularly non-standing-order transactions into savings accounts that are related to entropy, and it might thus be reasonable to exclude standing orders. However, while we cannot perfectly identify standing orders, they seem to account only for a small proportion of all savings account transfers and are thus unlikely to affect our results. Because of that, we do not exclude them.

Figure 3: Monthly flows in and out of savings accounts



Notes: Flows are calculated for each user-month as the total inflows, outflows, and flows (calculated as inflows - outflows) into all of a users's savings accounts. Zero net flows represent months where inflows are either perfectly balanced by outflows or where there were no flows at all.

2.5 Independent variable

Spending entropy:

- We calculate spending entropy using the Shannon entropy H(Shannon 1948), defined as

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

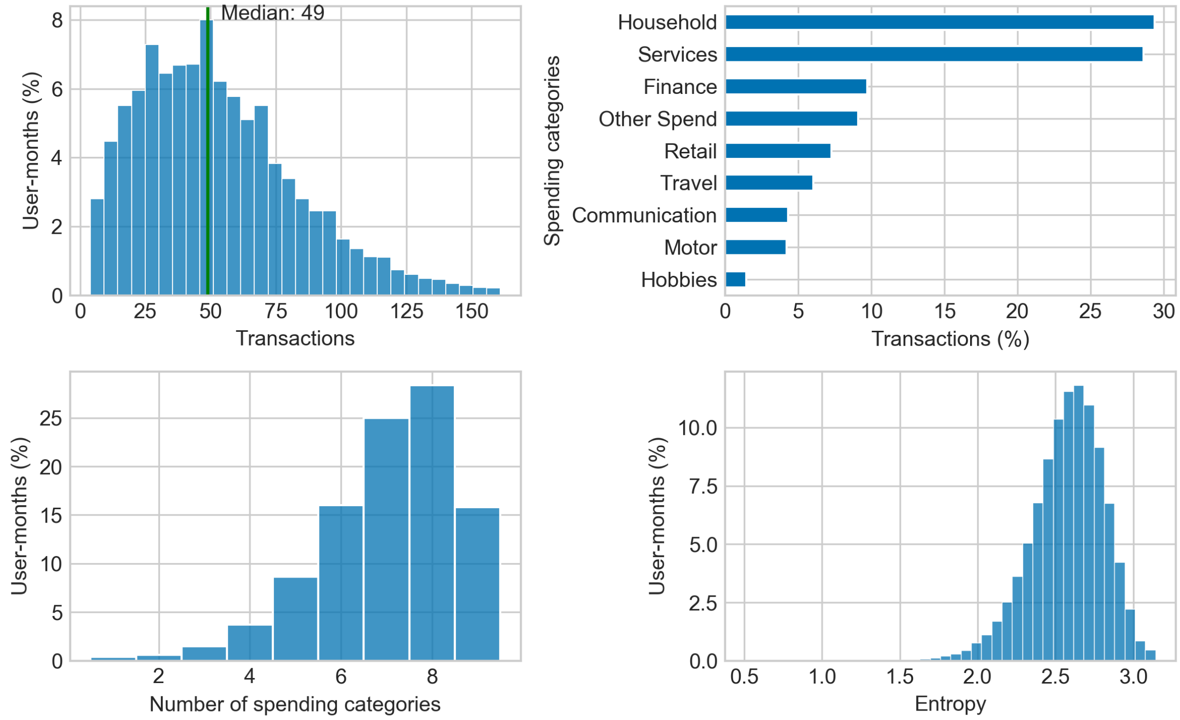
  where p_i is the probability that an individual makes a purchase in spending category i, and log is the base 2 logarithm. The measure can broadly be interpreted as the degree to which an individual's spending pattern is predictable, whith a higher score indicating less predictability.

- To calculate individual 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). Transactions included in the calculation are those classified as one of those spending categories that are debits and were made either from an indivuals current or credit card account.

- Also following that paper, when calculating p_i we use additive smoothing and add one to the numerator and N_{SC} to the denominator to avoid taking logs of zero counts in cases where an individual makes no purchases in a given spending category. p_i is thus calculated as

$$p_i = \frac{\text{Count of purchases in } SC_i + 1}{\text{Count of all purchases} + 9} \tag{2}$$

Figure 4: 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.

2.6 Control variables

We calculate age as an individual's approximate age at the time of the transactions, by subtracting a user's year of birth from the year the transaction took place.

3 Methods

3.1 Model specification

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

$s_{i,t}$ is individual i's savings rate in month t, calculated as the total inflow of funds in month t into all savings accounts held by i, divided by i's estimated monthly income.

The vector of control variables, $X_{i,t}$, contains the monthly spend for each spending category, total monthly spend across all categories, and annual income.

4 Results

Table 3: Main results

	has_sa_inflows			
	(1)	(2)	(3)	(4)
entropyz	−0.023***	−0.019***	−0.018***	−0.018***
	(0.002)	(0.002)	(0.002)	(0.002)
month_spend		0.00002***	0.00001***	
		(0.00000)	(0.00000)	
month_income			0.00002***	0.00002***
			(0.00000)	(0.00000)
spend_communication				0.0001***
				(0.00003)
spend_services				0.00002***
				(0.00000)
spend_finance				0.00002***
				(0.00000)
spend_motor				0.0001***
				(0.00002)
spend_travel				0.00003***
				(0.00000)
spend_hobbies				0.0002***
				(0.00003)
spend_other_spend				0.00002***
				(0.00000)
spend_household				0.00001***
				(0.00000)
spend_retail				0.0001***
				(0.00001)
month2	−0.003	0.003	0.003	0.005
	(0.006)	(0.006)	(0.006)	(0.006)
month3	−0.006	−0.003	−0.006	−0.006
	(0.006)	(0.006)	(0.006)	(0.006)
month4	−0.021***	−0.018***	−0.018***	−0.018***
	(0.006)	(0.006)	(0.006)	(0.006)
month5	−0.020***	−0.018***	−0.017***	−0.018***
	(0.006)	(0.006)	(0.006)	(0.006)
month6	−0.024***	−0.021***	−0.021***	−0.021***
	(0.006)	(0.006)	(0.006)	(0.006)
month7	−0.012*	−0.011*	−0.010	−0.013**
	(0.006)	(0.006)	(0.006)	(0.006)
month8	−0.004	−0.002	−0.002	−0.005
	(0.006)	(0.006)	(0.006)	(0.006)
month9	−0.010	−0.007	−0.007	−0.008
	(0.006)	(0.006)	(0.006)	(0.006)
month10	−0.001	−0.0003	−0.0003	−0.004
	(0.006)	(0.006)	(0.006)	(0.006)
month11	−0.015**	−0.013**	−0.013**	−0.016**
	(0.006)	(0.006)	(0.006)	(0.006)
month12	−0.003	−0.005	−0.006	−0.011*
	(0.006)	(0.006)	(0.006)	(0.006)
N	85,364	85,364	85,364	85,364
R^2	0.002	0.007	0.012	0.014
Adjusted R^2	−0.032	−0.027	−0.022	−0.020

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.

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*.

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

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.

Hacioglu, Sinem, Diego Känzig, and Paolo Surico (2020). "The Distributional Impact of the Pandemic". In:

MAS, Money Advice Service (2014). *Money Lives: the financial behaviour of the UK*. URL: https://www.moneyadviceservice.org.uk/en/corporate/money-lives (visited on 04/07/2020).

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.