Entropy[*]

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

February 15, 2022

Contents

1 Introduction

Story:

- This paper tests whether people save less in times when their lives are more chaotic. We define savings as ... and capture chaotic lives using ...
- Many people in UK and US struggle to save enough.
- This is well documented and thoroughly studied for retirement savings.
- But these are not only savings. People in UK and US also don't have enough to cover unexpected outlays.

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

- This could have important consequences: 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 (scarcity research)
- This paper aims to start to fill gap and study short-term savings behaviour.
- In doing so, we aim to contribute to three literatures:
  - Understanding savings behaivour
  - Understanding effect of entropy (muggleton)
  - Use of high-frequency transaction data (itself a sub-literature of use of newly available large-scale datasets)

2 Method

2.1 Dataset description

We use data from Money Dashboard (MDB), a financial management app that allows its users to link accounts from different banks to obtain an integrated view of their finances.[1] 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.[2]

2.2 Preprocessing and sample selection

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

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

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

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 Variable description

Outcome variable: Our outcome variable is a binary indicator for whether or not a user has made any payments into their savings accounts in a given month. 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. 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.

Variable of interest: 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:[3]

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

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

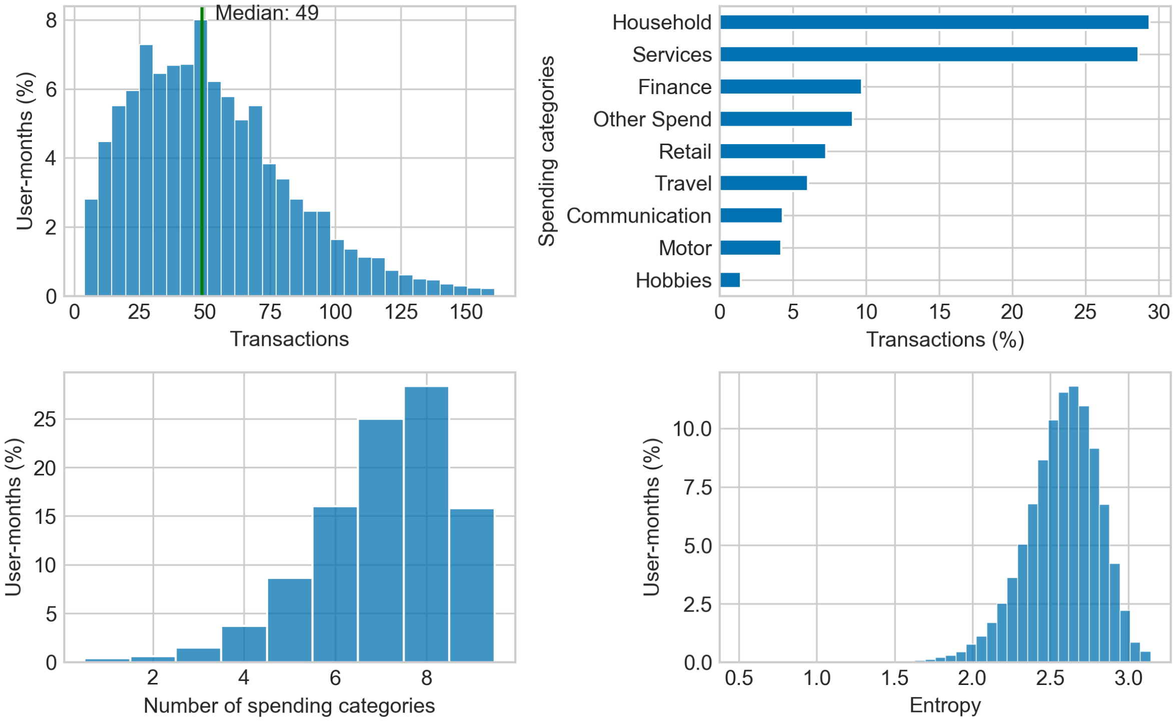
[3] Shannon entropy is customarily denoted as H following Shannon's own naming after Ludwig Boltzman's 1872 H-theorem in statistical mechanics, to which it is analagous.

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

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

Figure 1 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 1: 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.

2.4 Summary statistics

2.5 Model specification

We estimate models of the form:

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

where $I(s)_{i,t}$ 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. $X_{i,t}$ is a vector of control variables.

Table 2: Main results

	Dependent variable: has transfers into savings account			
	(1)	(2)	(3)	(4)
Entropy	-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 household				0.00001^{***}
				(0.00000)
Spend retail				0.0001^{***}
				(0.00001)
Spend other				0.00002^{***}
				(0.00000)
Individual fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Observations	85,364	85,364	85,364	85,364
R^2	0.002	0.007	0.012	0.014

Note: ... $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

3 Results

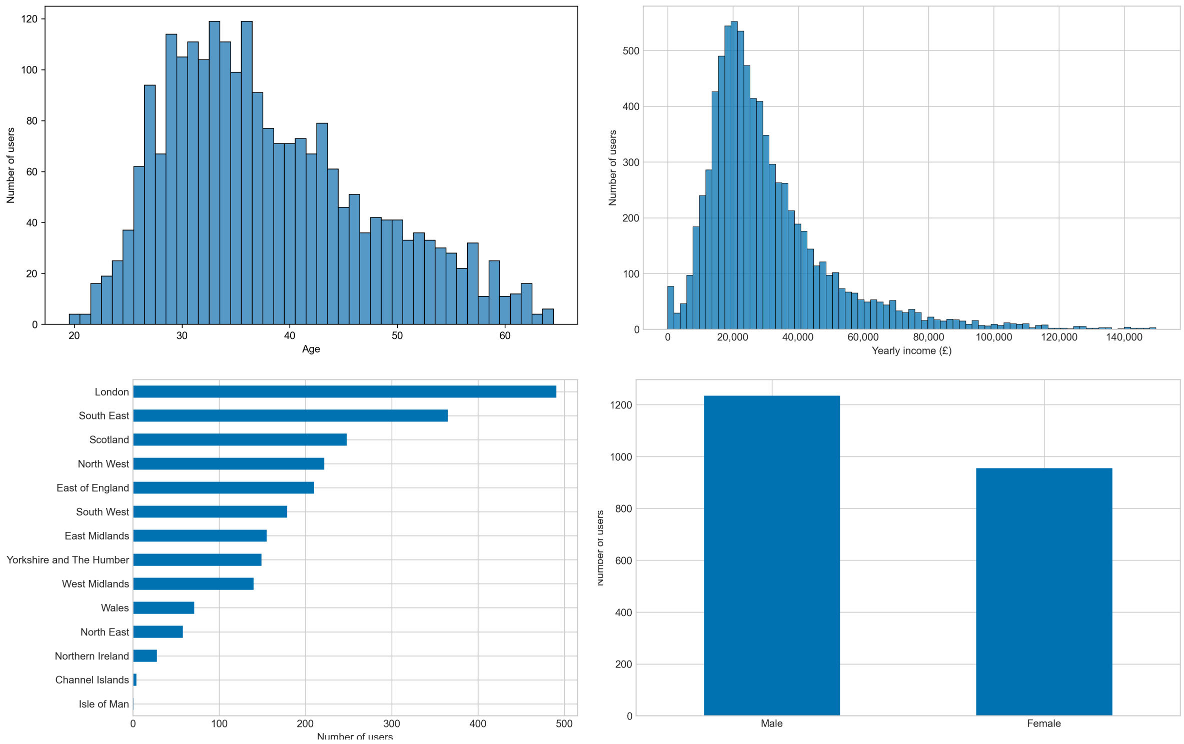
4 Discussion

Areas of further study:

- Does entropy as defined here really capture behaviour we're interested in?

A Data

Figure 2: Demographic characteristics of Money Dashboard users



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.

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.

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

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.

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.

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: