Entropy*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

January 5, 2022

Contents

1 Introduction

Nomenclature:

- user : Individual - ‘tag‘ : Spending categories

Literature:

Muggleton et al. (2020) find that consumption entropy over categories correlates with financial distress.

Davenport et al. (2020) study the impact of COVID-19 on the spending and savings behaviour of MDB users.

Baker and Kueng (2021) summarises literature that uses mass financial transaction data to study household financial behaviour.

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

Becker (2017) finds that access to a fintech money management app increases first-time savings and savings account balances among 65,000 customers of a large European bank but that update is negatively correlated with financial sophistication.

Colby and Chapman (2013) has useful literature review on short-term savings and suggests that subgoals can increase willingness to forego short-amounts in the present because they move the reference point in a prospect-theory framework.

Paper:

Independent variable: entropy over categories and others

Outcome variables: first-time saving, average monthly savings

2 Data

2.1 Preprocessing

Duplicate transactions

2.2 Sample selection

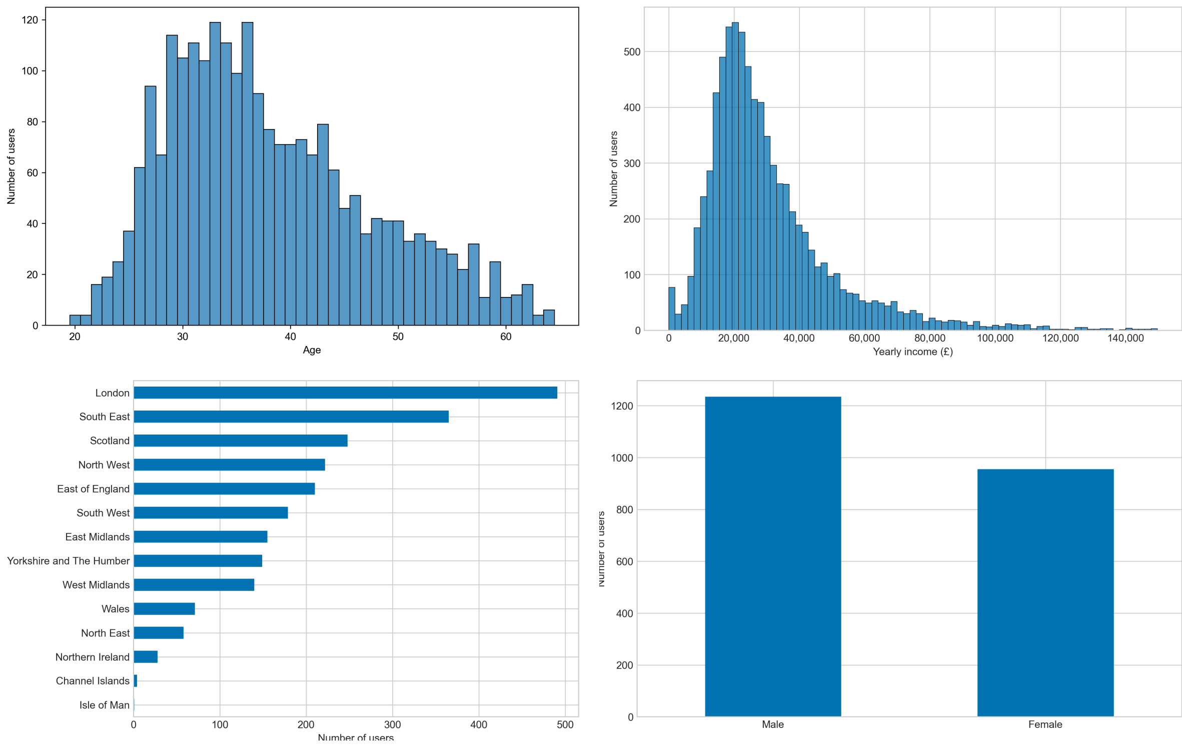
Table 1: Sample selection

	Users	Accounts	Transactions	Value (£M)
Raw sample	24,163	123,625	59,647,019	11,209.7
At least 6 months of data	21,508	117,000	59,088,076	11,116.8
No missing months	18,550	98,513	51,252,412	9,606.4
Account balances available	14,714	85,590	44,163,629	8,618.7
At least 5 debits totalling £200 per month	12,296	70,981	38,618,825	7,469.7
At least one current account	12,148	70,366	38,280,116	7,421.9
Income in 2/3 of all observed months	9,963	60,151	33,270,455	6,490.2
Yearly income between £5k and £200k	6,716	36,829	21,460,727	3,677.9
No more than 10 accounts in any year	6,169	26,999	18,241,355	2,855.1
Debits of less than £100k each month	5,760	24,486	16,409,561	1,998.7
Final sample	5,760	24,486	16,409,561	1,998.7

2.3 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',

Figure 1: Demographic characteristics of Money Dashboard users



'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).

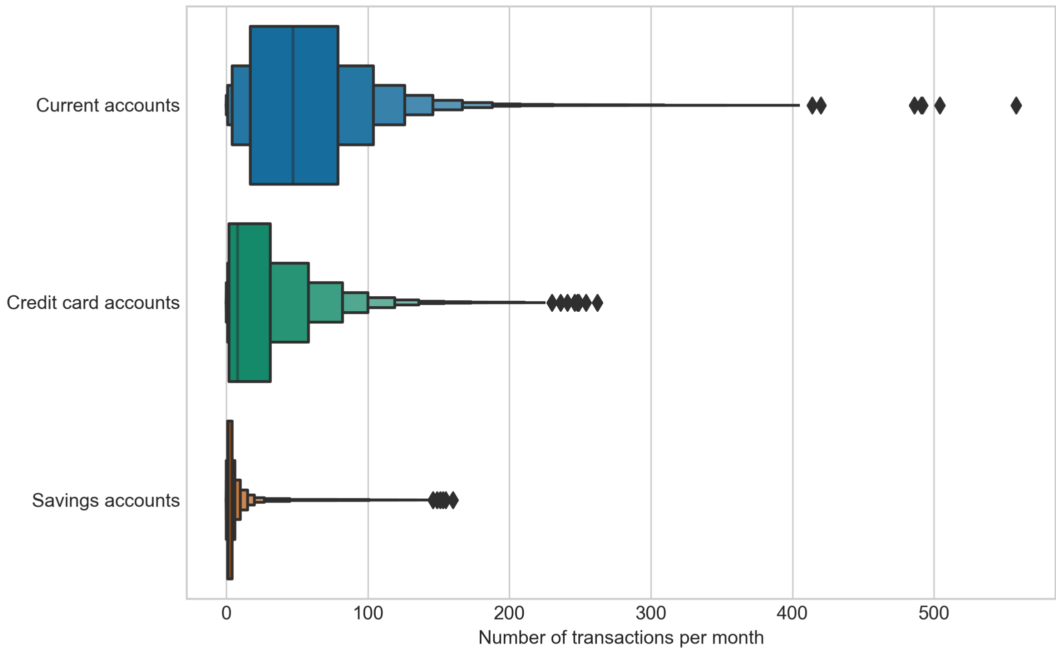
The main advantages of the data for the study of consumer financial behaviour are its high (transaction-level) frequency, that it is automatically collected and updated and thus less prone to errors and unaffected by biases that bedevil survey measures, and that it offers a view of consumers' entire 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), a fact that is reflected in our data. 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] 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

[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.

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.

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.

2.4 Dependent variable

- Add notes on excluding non-standing orders from ipynb.

Types of savings:

- Current account balances
- Savigns account balances
- ISA

Table 2: Summary statistics

	count	mean	std	min	max	25%	50%
obs	172845	94.938	52.1099	1	803	59	86
balance_ca	170936	1750.21	10200.3	-564218	208641	-300.42	393.829
balance_sa	68867	3153.2	10003.3	-209248	307243	6.67334	450.25
sa_inflows	63565	1025.29	3292.99	0	125040	60	251
sa_outflows	63565	945.789	2762.89	0	78773	0	200
sa_net_inflows	63565	79.4961	3311.73	-69750	120000	-177.48	19.86
sa_scaled_inflows	63565	53.8177	206.707	0	11614.7	3.27299	13.8856
sa_scaled_outflows	63565	50.388	173.423	0	9694.17	0	10.8786
sa_scaled_net_inflows	63565	3.42975	199.874	-9659.55	11614.7	-8.95894	0.897252
entropy_sptac	172845	2.57501	0.237914	0.52119	3.14224	2.4347	2.59672
total_monthly_spend	172845	1832.49	2549.84	-129296	90267.3	807.81	1394.2
tag_spend_retail	172845	66.4822	781.141	-62454	24719.8	5.34	50.38
tag_spend_services	172845	350.322	841.497	-32524.5	79621.2	109.5	234.81
tag_spend_household	172845	725.401	1382.33	-53325.4	85474.4	173.72	449.11
tag_spend_other_spend	172845	161.2	1118.75	-124874	80170	20.99	103.49
tag_spend_hobbies	172845	22.8894	146.678	-9277	20000	0	0
tag_spend_finance	172845	262.742	888.735	-67495.6	50757.1	7.51	70.48
tag_spend_travel	172845	122.63	577.963	-19765.1	67480	0	20.12
tag_spend_communication	172845	56.4601	117.003	-23697.6	7047.85	18.78	45
tag_spend_motor	172845	64.3664	190.035	-30371.4	18556.4	0	34.03

Types of balances, from Becker (2017), who treats balance at end of each month as observations:

- Current account balance
- Debit balance (savings and current account balance)
- Pure savings (savings account balance only)
- Credit balance (loans and negative current account)
- Pure credit (loans only)
- Wealth held (debit - credit balance)

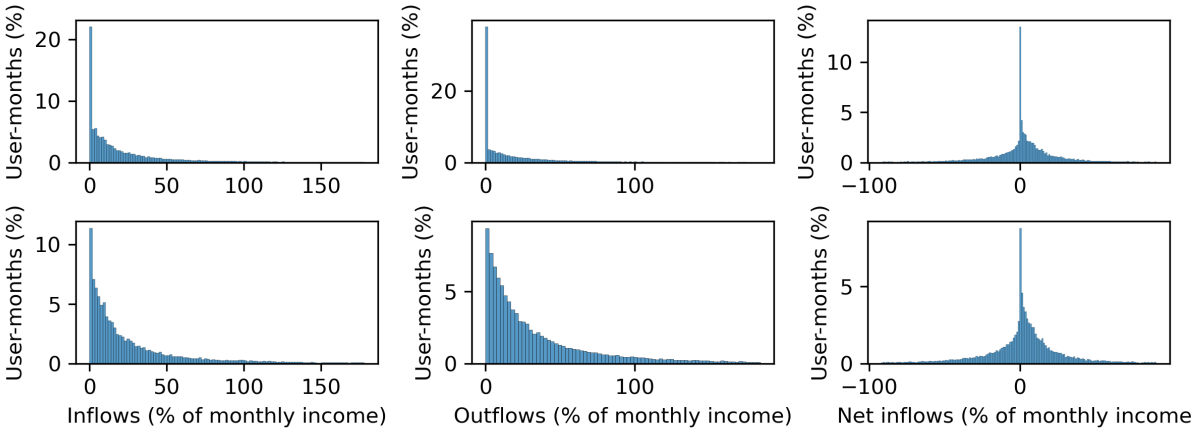
2.5 Independent variable

Spending entropy:

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

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

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.

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).

- 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

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

3 Methods

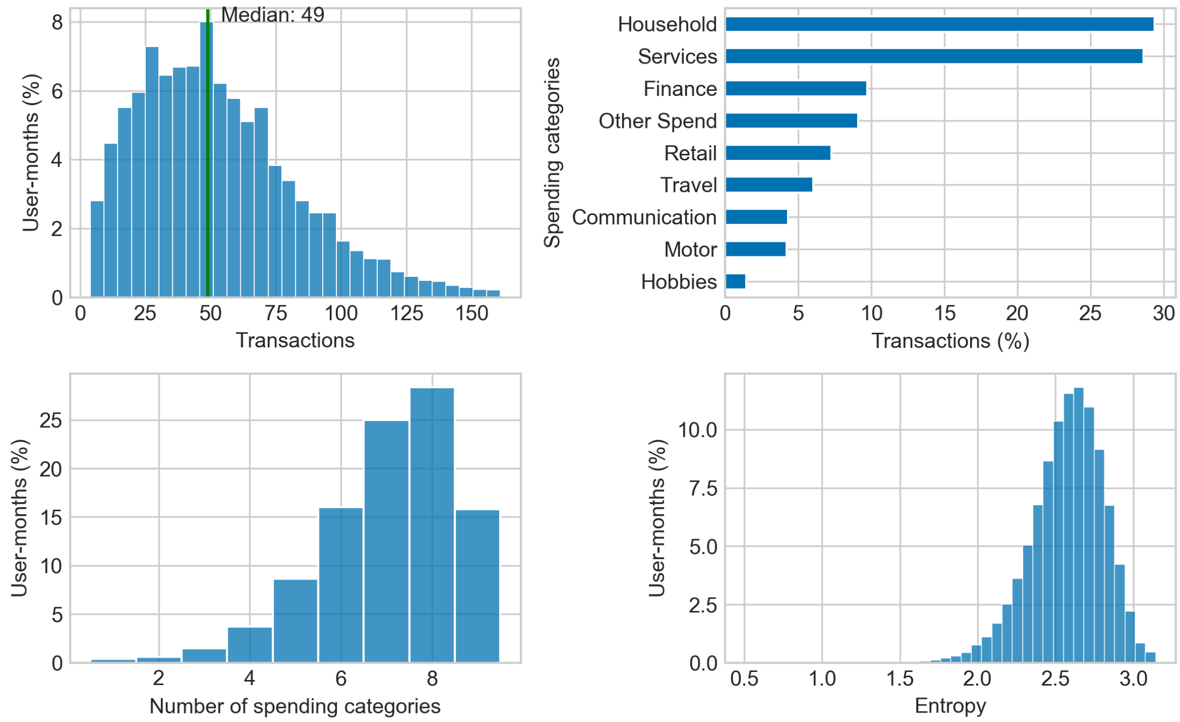
3.1 Model specification

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

$s_{i,t}$ 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.

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.

References

Baker, Scott R and Lorenz Kueng (2021). “Household Financial Transaction Data”. Tech. rep. National Bureau of Economic Research.

Becker, G (2017). “Does fintech affect household saving behavior? findings from a natural field experiment”. Tech. rep. mimeo.

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.

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

Davenport, Alex, Robert Joyce, Imran Rasul, and Tom Waters (2020). “Spending and saving during the COVID-19 crisis: evidence from bank account data”. In: *Institute for Fiscal Studies, Briefing Note* 308.

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.