Entropy*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

November 24, 2021

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.

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

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

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	27,129	132,900	65,871,348	12,194.8
At least 6 months of data	23,878	125,724	65,300,510	12,102.4
At least one current account	22,547	122,158	63,247,036	11,829.8
At least 5 monthly debits totalling GBP200	14,918	79,516	46,440,907	8,681.2
Income payments in 2/3 of all observed months	10,909	61,431	35,593,720	6,736.5
Yearly incomes between 5k and 200k	5,792	30,794	18,704,936	3,278.3
No more than 10 active accounts in any year	5,359	23,151	15,954,877	2,541.3
Debits of no more than 100k in any month	5,004	20,963	14,297,483	1,765.7
Current and savings account balances available	2,716	10,028	7,531,905	956.9
Last account refresh within observed period	2,716	10,021	7,531,838	956.9
Working-age	2,357	8,607	6,792,133	847.3
Final sample	2,357	8,607	6,792,133	847.3

2.3 Sample description

2.4 Dependent variable

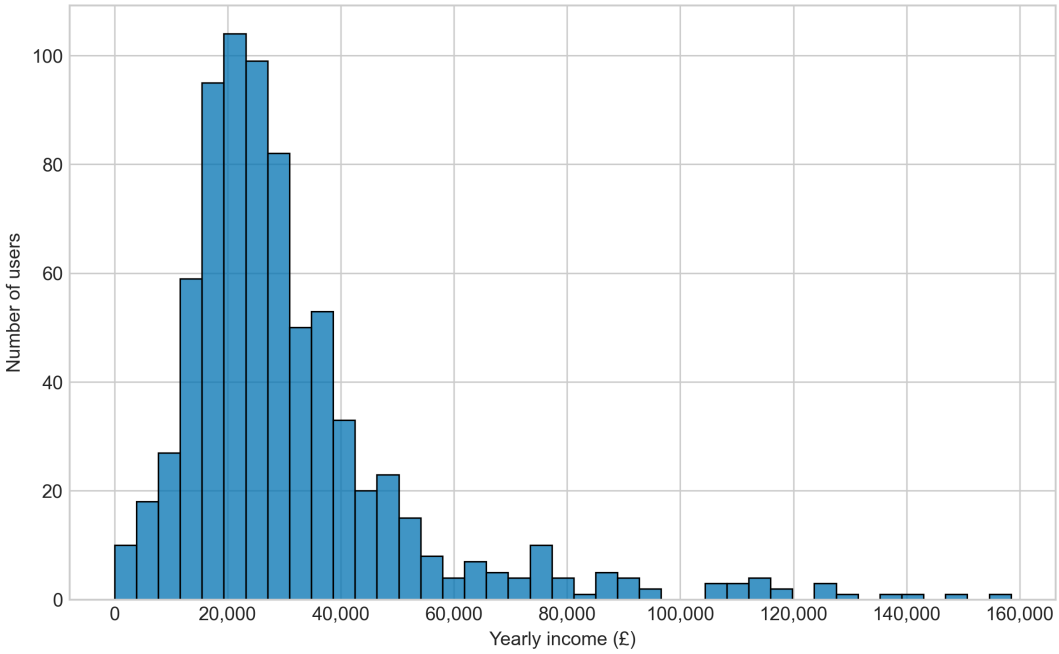
Types of balances, from Becker (2017), who treats balance at end of each month as observations:

- 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:

Figure 1: Distribution of user incomes



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

$$H = -\sum p_i log_2(p_i), \tag{1}$$

where p_i is the probability that an individual makes a purchase in spending category i. 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.

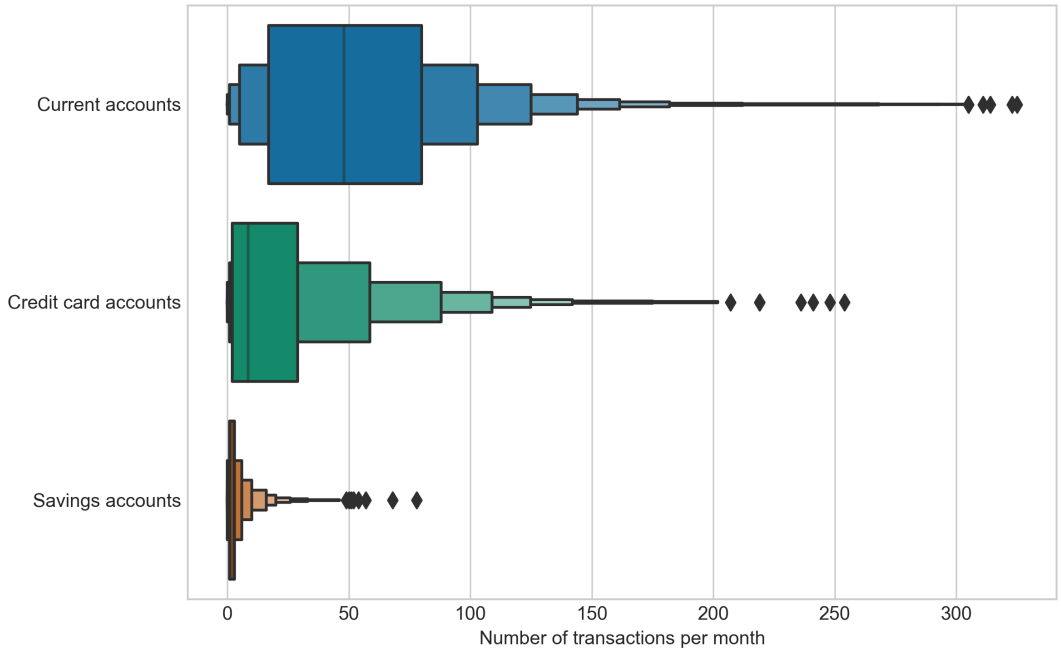
- To calculate individual entropy scores, we group spending into 9 spending categories (SC), based on the classification used by Lloyds Banking Group as discussed in Muggleton et al. (2020).

- 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

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

- The right-hand panel in Figure 3 shows the distribution of the resulting entropy scores.

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.

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.

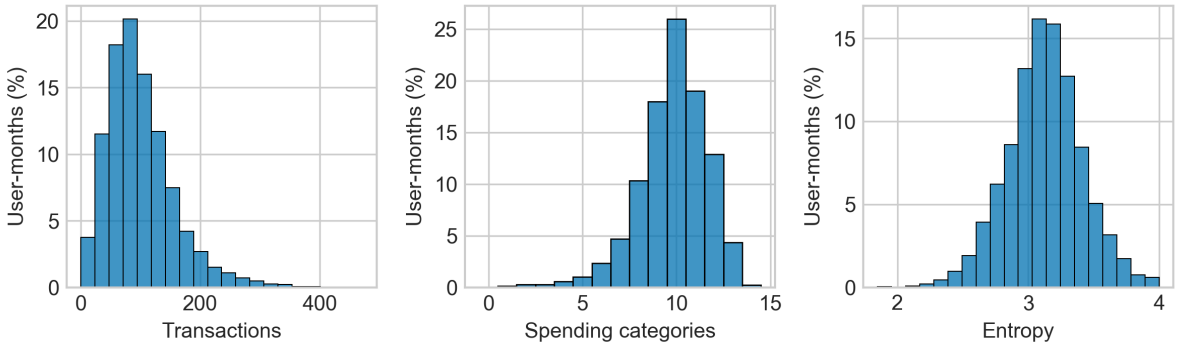
Colby, Helen and Gretchen B Chapman (2013). "Savings, subgoals, and reference points". In:

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.

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.

Figure 3: Transactions distributions



Notes: Distribution of the number of transactions per user-month (left), the number of different spending categories these transactions fall into (middle), and user-level entropy scores based on these same spending categories (right).