# Entropy[*]

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

February 14, 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.

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

- We capture the degree of chaos in a person's life using entropy.

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 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

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.

Our variable of interest is the degree of chaos of an individual's lifestyle. We ...

–todo–

### 2.4 Independent variable

Spending entropy:

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

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

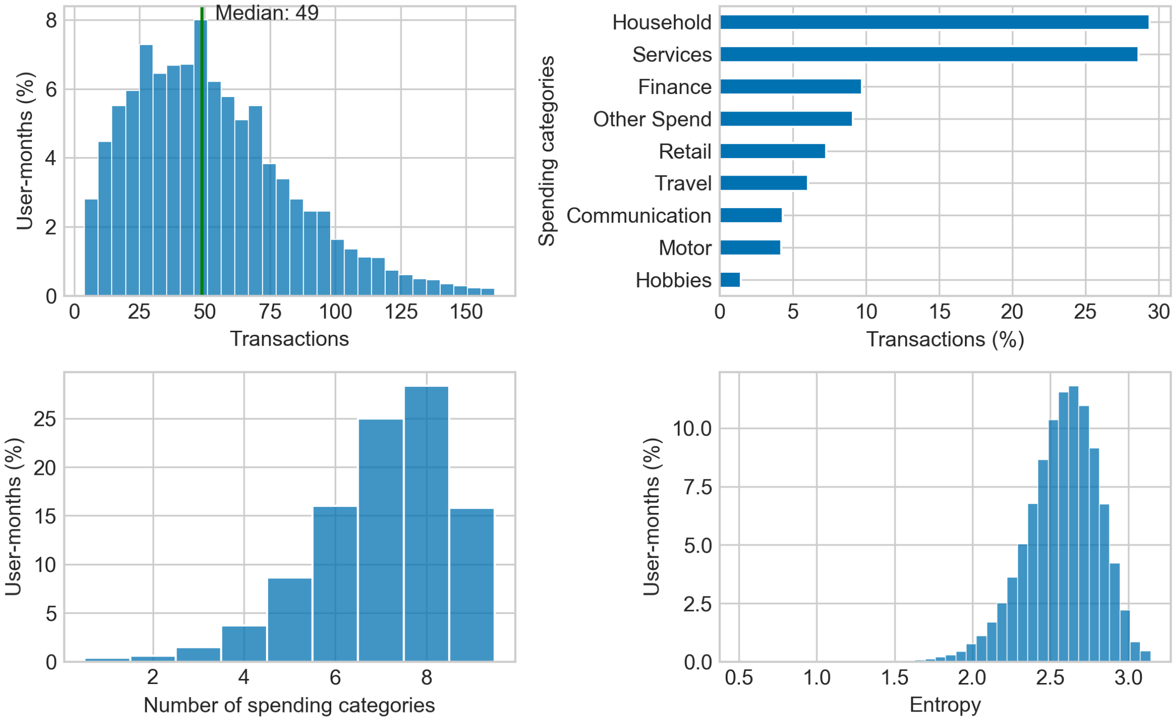
  where $p_i$ is the probability that an individual makes a purchase in spending category $i$, and $log$ is the base 2 logarithm. The measure can broadly be interpreted as the degree to which an individual's spending pattern is predictable, whith a higher score indicating less predictability.

- 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 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.

## 2.5 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.

## 2.6 Summary statistics

The four panels in Figure 2 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).[3]

[3] 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 2: Demographic characteristics of Money Dashboard users

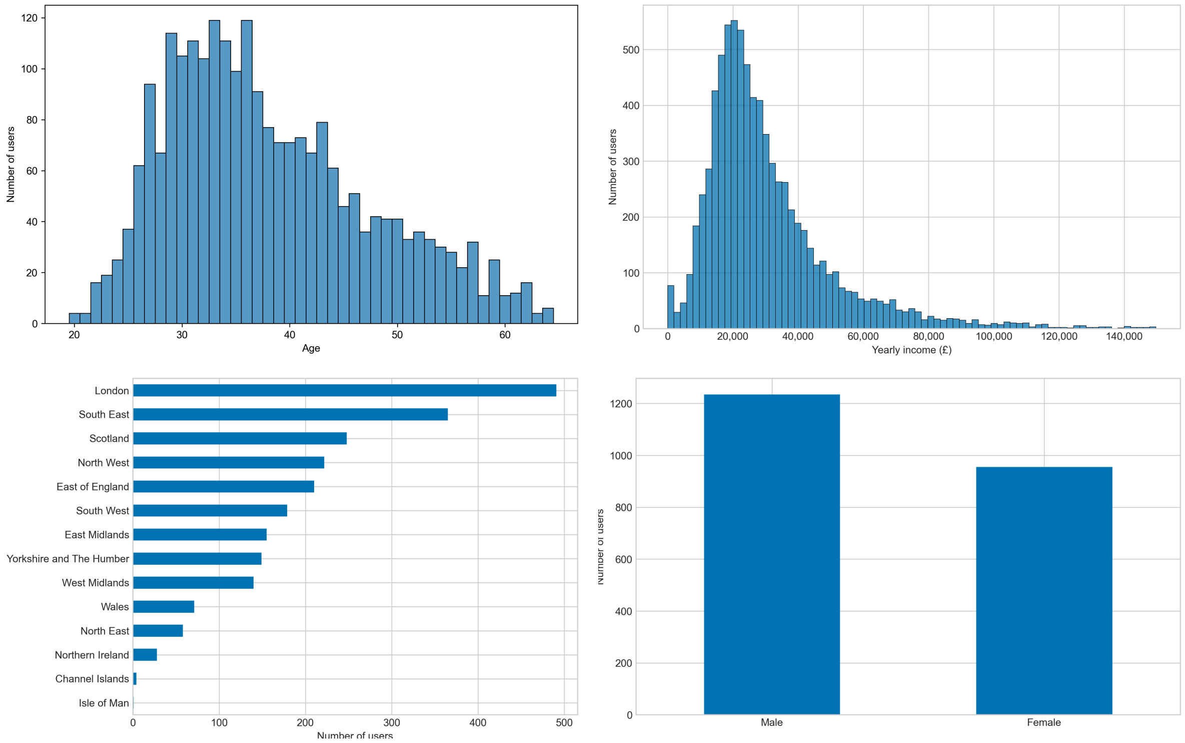


Table 2: Summary statistics

| | 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 | |

### 2.7 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.

## 3 Results

## 4 Discussion

Areas of further study:

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

## A Data

Data limitations:

Table 3: 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.

- 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.

# 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:

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.