# Entropy*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

March 11, 2022

**Abstract**

...

- This paper aims to better understand the drivers of emergency savings.
  - We test the important of commonly discussed determinants (financial behaviour, planning, characteristics), see reports.
  - Entropy has recently been found to affect financial outcomes. We test whether entropy has impact above and beyond traditional factors, or can be thought of as a summary statistic of (some) of them.
- Large literature on savings, but academic literature almost exclusively focused on pensions.
- Short-term savings matter, too.
- People in UK and US also don't have enough to cover unexpected outlays. (See reports)
- This has important consequences:
  - Short-term: financial well-being (see reports)
  - Long-term (viscious cycle): 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 (less savings leading to increased risk of financial hardship leading to more stress leading to less savings...)
- This paper aims to start to fill gap and study short-term savings.
- Main findings:
  - Chaotic lifestyle impairs saving above and beyond total amount spend and income earned, implying that lack of saving is not only due to incomes too low to save.
  - What is entropy? Convenient measure to pick up stress in person's life?

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

  - Mechanism: scarcity.

- In doing so, we aim to contribute to three literatures:
  - Main 1: Understanding emergency savings behaivour (nest, aspen reports)
  - Main 2: Understanding effect of behavioural entropy - eliciting useful personality characteristics from large-scale data
  - Also 1: More broadly: part of savings literature (pension literature, savings buffer)
  - Also 2: Use of high-frequency transaction data (itself a sub-literature of use of newly available large-scale datasets)

# 1 Results

## 1.1 Use-month level analysis

In this section, we analyse results at the user-month level. Entropy measures are calculated based on the probability of occurrence of a pattern within a user month. Implicitly, this assumes that move diverse spending reflects more chaotic behaviour.

The outcome variable is a dummy indicating whether a user made at least one transfer into their savings account in a given month.

Table 1 shows controls added one at a time.

Table 1: Controls included one-by-one

| Dependent Variable: | has_sa_inflows | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| *Variables* | | | | | | | | | | | |
| entropy_tag | 0.0329*** | 0.0341*** | 0.0322*** | 0.0230** | 0.0329*** | 0.0263** | 0.0293*** | 0.0360*** | 0.0305*** | 0.0316*** | 0.0329*** |
| | (0.0110) | (0.0101) | (0.0109) | (0.0108) | (0.0110) | (0.0108) | (0.0109) | (0.0109) | (0.0109) | (0.0110) | (0.0109) |
| Regular savings | | 0.2376*** | | | | | | | | | |
| | | (0.0080) | | | | | | | | | |
| prop_credit | | | 0.1269*** | | | | | | | | |
| | | | (0.0244) | | | | | | | | |
| month_spend | | | | 0.0162*** | | | | | | | |
| | | | | (0.0013) | | | | | | | |
| is_urban | | | | | -0.1472 | | | | | | |
| | | | | | (0.1799) | | | | | | |
| month_income | | | | | | 0.0263*** | | | | | |
| | | | | | | (0.0025) | | | | | |
| year_income | | | | | | | 0.0033*** | | | | |
| | | | | | | | (0.0005) | | | | |
| has_regular_income | | | | | | | | 0.0488*** | | | |
| | | | | | | | | (0.0070) | | | |
| has_month_income | | | | | | | | | 0.1263*** | | |
| | | | | | | | | | (0.0177) | | |
| loan_repmt | | | | | | | | | | 0.0192* | |
| | | | | | | | | | | (0.0109) | |
| pdloan_repmt | | | | | | | | | | | 0.0004 |
| | | | | | | | | | | | (0.0230) |
| *Fixed-effects* | | | | | | | | | | | |
| user_id | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| month | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | | | | | | |
| Observations | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 |
| $R^2$ | 0.44276 | 0.46509 | 0.44370 | 0.44561 | 0.44287 | 0.44634 | 0.44520 | 0.44449 | 0.44356 | 0.44290 | 0.44276 |
| Within $R^2$ | 0.00035 | 0.04040 | 0.00204 | 0.00546 | 0.00054 | 0.00677 | 0.00473 | 0.00346 | 0.00177 | 0.00059 | 0.00035 |

*Clustered (user_id) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2 summarises shows controls added cumulatively.

Table 2: Controls included cumulatively

| Dependent Variable: | has_sa_inflows | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| *Variables* | | | | | | | | | | | |
| entropy_tag | 0.0329*** | 0.0341*** | 0.0337*** | 0.0256** | 0.0256** | 0.0217** | 0.0216** | 0.0185* | 0.0175* | 0.0173* | 0.0173* |
| | (0.0110) | (0.0101) | (0.0101) | (0.0100) | (0.0100) | (0.0100) | (0.0100) | (0.0099) | (0.0099) | (0.0099) | (0.0099) |
| Regular savings | | 0.2376*** | 0.2351*** | 0.2316*** | 0.2313*** | 0.2283*** | 0.2276*** | 0.2506*** | 0.2514*** | 0.2514*** | 0.2514*** |
| | | (0.0080) | (0.0080) | (0.0080) | (0.0079) | (0.0079) | (0.0080) | (0.0087) | (0.0087) | (0.0087) | (0.0087) |
| prop_credit | | | 0.0690*** | 0.0543** | 0.0543** | 0.0545** | 0.0538** | 0.0676*** | 0.0678*** | 0.0682*** | 0.0682*** |
| | | | (0.0224) | (0.0224) | (0.0224) | (0.0223) | (0.0223) | (0.0221) | (0.0221) | (0.0220) | (0.0220) |
| month_spend | | | | 0.0132*** | 0.0132*** | 0.0111*** | 0.0110*** | 0.0113*** | 0.0113*** | 0.0112*** | 0.0113*** |
| | | | | (0.0012) | (0.0012) | (0.0011) | (0.0011) | (0.0011) | (0.0011) | (0.0011) | (0.0011) |
| is_urban | | | | | -0.0865 | -0.1013 | -0.1030 | -0.1034 | -0.1045 | -0.1044 | -0.1044 |
| | | | | | (0.1466) | (0.1537) | (0.1552) | (0.1541) | (0.1559) | (0.1559) | (0.1559) |
| month_income | | | | | | 0.0208*** | 0.0187*** | 0.0189*** | 0.0166*** | 0.0166*** | 0.0166*** |
| | | | | | | (0.0022) | (0.0018) | (0.0018) | (0.0019) | (0.0019) | (0.0019) |
| year_income | | | | | | | 0.0006 | 0.0008* | 0.0008* | 0.0008* | 0.0008* |
| | | | | | | | (0.0004) | (0.0004) | (0.0004) | (0.0004) | (0.0004) |
| has_regular_income | | | | | | | | -0.0420*** | -0.0430*** | -0.0432*** | -0.0432*** |
| | | | | | | | | (0.0077) | (0.0078) | (0.0078) | (0.0078) |
| has_month_income | | | | | | | | | 0.0723*** | 0.0725*** | 0.0725*** |
| | | | | | | | | | (0.0168) | (0.0168) | (0.0168) |
| loan_repmt | | | | | | | | | | 0.0040 | 0.0040 |
| | | | | | | | | | | (0.0095) | (0.0095) |
| pdloan_repmt | | | | | | | | | | | -0.0028 |
| | | | | | | | | | | | (0.0204) |
| *Fixed-effects* | | | | | | | | | | | |
| user_id | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| month | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *Fit statistics* | | | | | | | | | | | |
| Observations | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 | 82,589 |
| $R^2$ | 0.44276 | 0.46509 | 0.46536 | 0.46725 | 0.46729 | 0.46946 | 0.46951 | 0.47054 | 0.47077 | 0.47078 | 0.47078 |
| Within $R^2$ | 0.00035 | 0.04040 | 0.04089 | 0.04428 | 0.04434 | 0.04825 | 0.04833 | 0.05018 | 0.05060 | 0.05061 | 0.05061 |

*Clustered (user_id) standard-errors in parentheses*
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3 replicates results from Muggleton et al. (2020) Tables S20 (Columns 1 and 2) and Table S40 (columns 3 and 4).

## 2 Discussion

Areas of further study:

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

## 3 Methods

**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]

**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 4 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 3: Muggleton et al. (2020) replication

| Dependent Variable: | id | | | |
|---|---|---|---|---|
| Model: | (1) | (2) | (3) | (4) |
| *Variables* | | | | |
| Entropy (auto-tag-based, smooth) | -0.0703*** | | -0.0317*** | |
| | (0.0017) | | (0.0034) | |
| Entropy (auto-tag-based) | | 0.0466*** | | 0.0249*** |
| | | (0.0028) | | (0.0045) |
| Spend communication | 0.5852*** | 0.6391*** | 0.0997*** | 0.1148*** |
| | (0.0254) | (0.0258) | (0.0381) | (0.0381) |
| Spend finance | 0.0299*** | 0.0185*** | 0.0125*** | 0.0119*** |
| | (0.0028) | (0.0028) | (0.0036) | (0.0037) |
| Spend hobbies | -0.0388 | -0.0345 | 0.0435* | 0.0473* |
| | (0.0279) | (0.0282) | (0.0259) | (0.0261) |
| Spend household | -0.0119*** | -0.0035** | -0.0029 | 0.0012 |
| | (0.0016) | (0.0016) | (0.0021) | (0.0021) |
| Spend other | 0.0101** | 0.0186*** | -0.0069* | -0.0027 |
| | (0.0039) | (0.0040) | (0.0041) | (0.0040) |
| Spend motor | 0.0600*** | 0.0938*** | 0.0066 | 0.0440** |
| | (0.0152) | (0.0155) | (0.0210) | (0.0210) |
| Spend retail | -0.0376*** | -0.0061 | -0.0087 | -0.0020 |
| | (0.0074) | (0.0074) | (0.0069) | (0.0070) |
| Spend services | -0.0534*** | -0.0107*** | -0.0234*** | -0.0077** |
| | (0.0036) | (0.0034) | (0.0041) | (0.0037) |
| Spend travel | -0.0515*** | -0.0366*** | -0.0058 | -0.0024 |
| | (0.0048) | (0.0049) | (0.0036) | (0.0036) |
| Female | 0.0454*** | 0.0395*** | | |
| | (0.0030) | (0.0030) | | |
| Age | -0.0019*** | -0.0026*** | -0.0093*** | -0.0078*** |
| | (0.0001) | (0.0001) | (0.0023) | (0.0024) |
| Year income | -0.0016*** | -0.0015*** | 0.0005* | 0.0006** |
| | $(9.17 \times 10^{-5})$ | $(9.24 \times 10^{-5})$ | (0.0003) | (0.0003) |
| (Intercept) | 0.2913*** | 0.2810*** | | |
| | (0.0056) | (0.0056) | | |
| *Fixed-effects* | | | | |
| User id | | | Yes | Yes |
| Calendar month | | | Yes | Yes |
| *Fit statistics* | | | | |
| Observations | 82,589 | 82,589 | 82,589 | 82,589 |
| $R^2$ | 0.04435 | 0.02847 | 0.63465 | 0.63343 |
| Within $R^2$ | | | 0.00744 | 0.00414 |

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Sample selection

| | Users | User-months | Accounts | Transactions |
|---|---|---|---|---|
| Raw sample | 26,513 | 739,047 | 130,058 | 64,464,494 |
| Annual income of at least £10k | 8,154 | 199,274 | 37,248 | 20,381,666 |
| Income in 2/3 of all observed months | 8,032 | 197,422 | 36,805 | 20,213,950 |
| At least one savings account | 4,751 | 127,958 | 28,223 | 13,776,356 |
| At least 6 months of data | 4,283 | 126,703 | 26,749 | 13,668,274 |
| Monthly debits of at least £200 | 3,555 | 103,604 | 21,903 | 11,670,234 |
| Five or more monthly current account txns | 3,312 | 96,588 | 20,247 | 10,771,752 |
| Spends in two distinct categories each month | 3,276 | 95,350 | 19,973 | 10,668,163 |
| Complete demographic information | 2,730 | 82,589 | 16,796 | 9,266,079 |
| Final sample | 2,730 | 82,589 | 16,796 | 9,266,079 |

**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.

MPS (2018) finds that saving habit is often more important than amount saved. On individual level, has saving habit dummy as outcome. Could also use 12-month rolling window.

Additional outcomes:

- Has overdraft fees (to see whether we can also reverse muggleton2020evidence result).

Notes on measuring savings:

- We can think of total savings as the sum of long-term and short-term savings. Long-term savings are savings for retirement, either individually or through an employer-linked pension scheme. These kinds of savings, and especially savings through pension schemes, are well researched. In contrast, there is almost no research on short-term savings, which comprise savings for particular goals such as a new car, a holiday, or a wedding, and emergency savings to have a buffer for unexpected events. (See nest2021supporting for more on emergency savings)
- We aim to close this gap. In our data, we cannot distinguish between goal-oriented and emergency savings, so we focus on total short-term savings[3]

**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

[3]MDB allows users to create custom tags and some users use them to indicate the intended use for their savings transactions (e.g. "wedding", "holidays"). But only a very small number of transactions have such tags, and we do not pursue this further.

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:[4]

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

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

We normalise $H$ by $log(N_{SC})$, the entropy of completely random shopping behaviour, so that it takes value between 0 and 1.[5]

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:

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

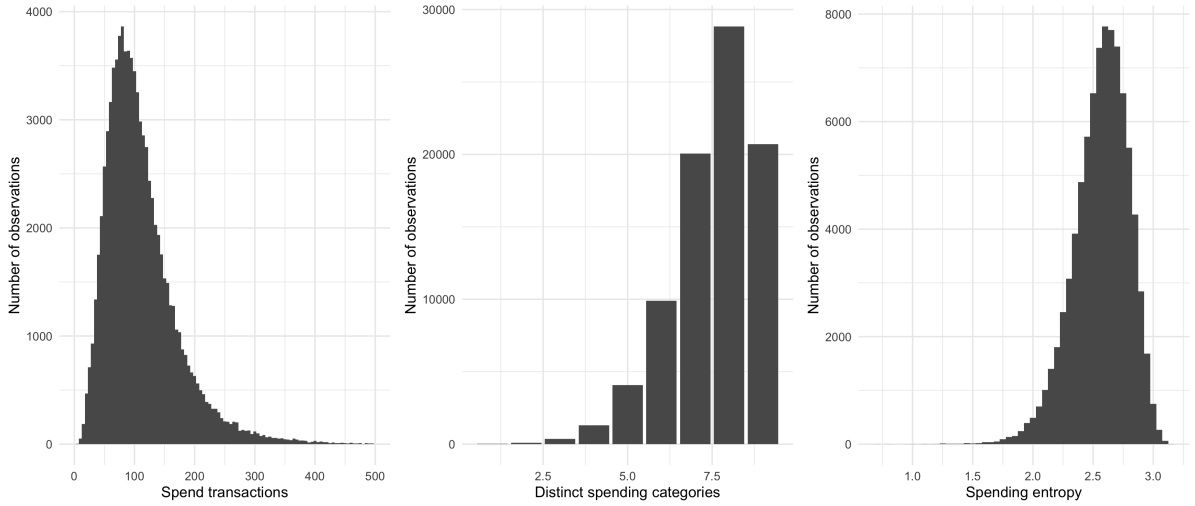
where $N_{SC}$ is the total number of categories.

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.

[4] 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.

[5] $log(N_{SC})$ is the probability of a completely random shopping pattern because for in this case, for $N_{SC}$ different spending categories, we would have $p_i = 1/N_{SC}$ for each category $i$ so that $H = -N_{SC}p_i log(p_i) = -log(p_i) = log(N_{SC})$.

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.

Issues from imperfect labelling of MDB data:

- Transaction tagging in the MDB data is imperfect: about 20 percent of transactions have no tag.
- This creates two issues for entropy calculation.
- First, entropy scores of high-entropy individuals are biased downwards - relatively more than those of low-entropy individuals. Reason: missing tags are not random: more common transactions such as groceries or take-away purchases are more likely to be tagged ...
- All zero count-cases: ... Solution: require minimum number of txns in two different labels (for all categories we use to calculate entropy). Two to avoid 0 entropy cases that are unlikely to be genuine but probably artefacts of missing labelling.

**Auto tag entropy**

- Use apriori algorithm
- minsup: minimum number of baskets a pattern is required to appear (else it's dropped)
- In our context: baseket is collection of auto tags with positive txsn counts in a user month, while pattern is pattern of such auto tags.
- Patterns also called 'representative baskets'.
- Algorithm steps (adapted from guidotti2015behavioural)
  - Identify all patterns (representative baskets)
  - Discard representative baskets that appear in fewer than minsup months we observe for a user.

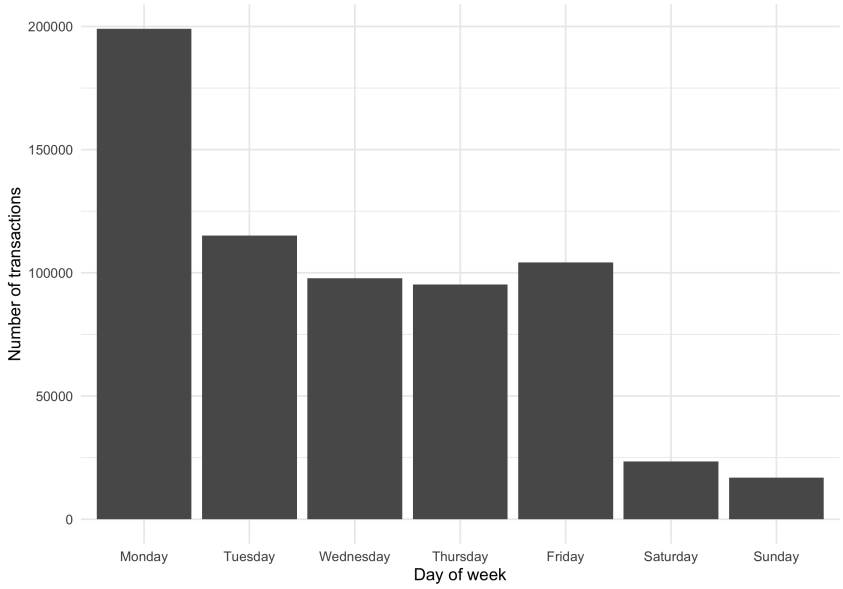
– Assign representative basked to each of the user's months.

– Calculate probabilities of observing a representative basked based on occurrences across all of a user's month. E.g. user with 5 months of data with representative baskets [1, 1, 2, 3, 4] has representative basket probabilitis 2/5 for repr basket 1, and 1/5 for repr baskets 2-4.

– Calculate user-leven entropy based on probabilities.

**Shopping-time based entropy** We calculate entropy based on the probability of $(dayofweek, merchant)$ tuples, where we follow Guidotti et al. (2015) and bin *day of week* into *weekends* and *weekday*, to reduce excessive fluctuations. Because banks tend to process weekend transactions on Monday, as shows in Figure 2, we cannot distinguish transactions made on Saturdays or Sundays from those made on Mondays, and thus classify all of them as weekend transactions.

Figure 2: Transactions by day of week




Notes: Number of transactions by day of week, based on a 1/1000 sample of the full data. Shows that banks process most weekend transactions on Mondays.

We drop the about 25 percent of transactions for which we cannot identify a merchant. The alternative would be leaving these transactions in the sample and treating "unknown merchant" as a single merchant. But for user-months for which the merchant is unknown for all transactions, this would lead to an entropy score of 0, which is undesireable.

**Summary statistics**

**Model specification** We estimate models of the form:

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

where $s_{i,t}$ 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, $H_{it}$ is $i$'s spending entropy in month

$t$, $x_{i,t}$ a vector of control variables, $\alpha_i$ an individual fixed effect, $\lambda_t$ a calendar month fixed effect, and $\epsilon_{i,t}$ the error term.

# 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*.

CAN, Commonwealth Bank of Australia (2019). "Improving the Financial Wellbeing of Australians". Tech. rep. URL: https://www.commbank.com.au/content/dam/caas/newsroom/docs/CWM0375_Financial%20Wellbeing%20Report_v4.pdf.

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

CFPB, Consumer Financial Protection Bureau (2017). "Financial Well-being in America". Tech. rep. URL: https://www.consumerfinance.gov/data-research/research-reports/financial-well-being-america/.

Finance, UK (2021). "Card Spending Update for November 2021". Tech. rep. URL: https://www.ukfinance.org.uk/sites/default/files/uploads/Data%20(XLS%20and%20PDF)/Card-Spending-Update-November-2021-final.pdf.

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.

Krumme, Coco, Alejandro Llorente, Manuel Cebrian, Alex Pentland, and Esteban Moro (2013). "The predictability of consumer visitation patterns". In: *Scientific reports* 3.1, pp. 1–5.

MPS, Money and Pension Service (2018). "Building the Financial Capability of UK Adults". Tech. rep. URL: https://moneyandpensionsservice.org.uk/2019/02/06/adult-financial-capability-building-the-financial-capability-of-uk-adults-survey/.

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:


# A Data

## A.1 Data preprocessing

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.

## A.2 Variable construction

We classify potential determinants of savings behaviour into *financial behaviours*, *financial planning*, and *individual or household characteristics*, a classification frequently used in policy research on the financial wellbeing (CAN 2019, CFPB 2017, MPS 2018).

Financial behaviour

- Regular savings, dummy for 10 out of last 12 months
- Proportion of purchases paid with credit card. This is only about 6 percent in our final sample, whereas it is 12 percent in the full sample.[6]
- Month total and category spend (category spend for robustness)

Planning

- Regular login, dummy for 1 / month in 10 out of last 12 months. Have login data for about 50 percent of sample, so best to work with full sample once I use it. Implement once I can do that. – not yet implemented –

Individual and household characteristics

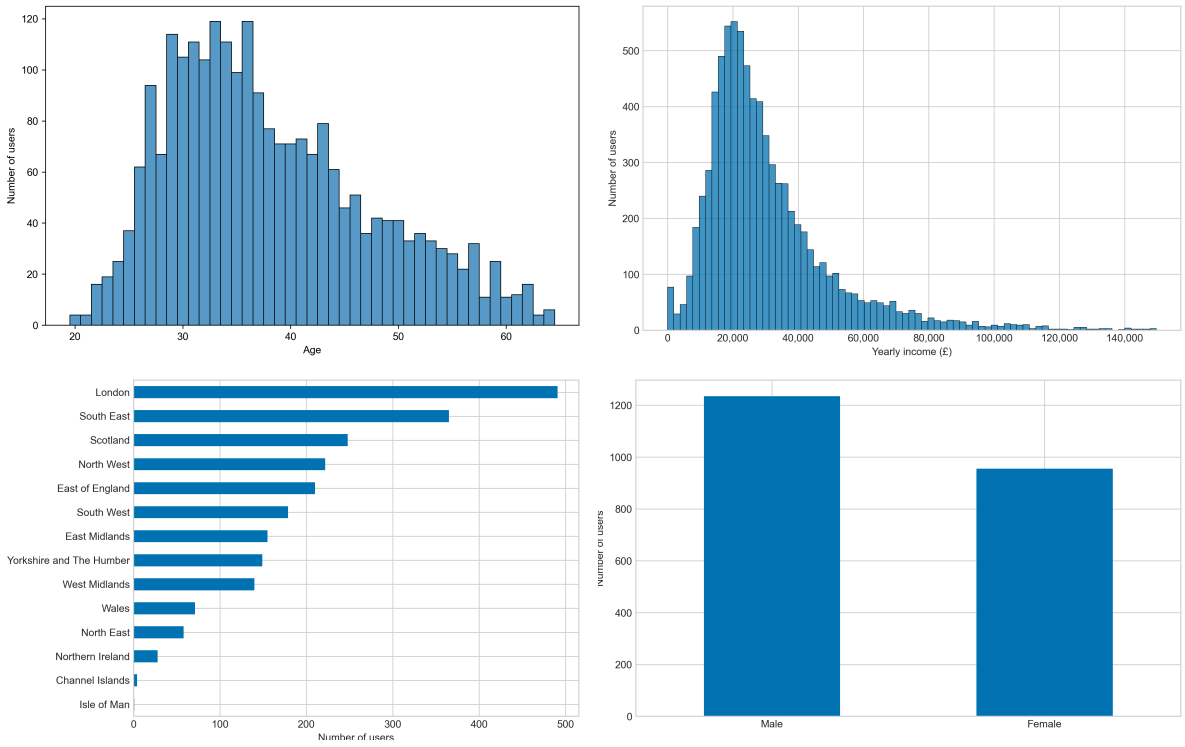
- Gender
- Age
- Urban
- Region
- Month income, winsorised at top 1 percent-level.
- Year income, winsorised at the 1 percent level.
- Regular income, dummy for 10 out of last 12 months
- Month income std – not implemented yet –

[6] Across the UK, the proportion of credit pard purchases is about 17 percent in a typical month (Finance 2021). The proportion in our data is likely lower because the sample is skewed towards more affluent individuals.

- Income current month, dummy for month income $> 0$
- Has children, imperfect – not yet implemented –
- Index of multiple deprivations from nspl – not implemented yet –
- Received benefits
- Receives pension
- Housing tenure: mortgage, rent, other (owning outright implied)
- Takes out (payday) loan
- Total balance or balance / avg. month spend – not yet implemented –

## A.3 Variabel description

Figure 3: Demographic characteristics of Money Dashboard users




Exploration of control variables here (e.g. like jpmorgan2019weathering for income stability)

## A.4 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.

# 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

## B.1 Entropy calculation

Entropy can be calculated along a number of dimensions.

- Category-based vs time-based vs category-time based (Guidotti et al. 2015, Krumme et al. 2013)
- Count-based vs value-based

- Intratemporal vs intertemporal (Krumme et al. 2013)
    - Based on behaviour within a given time period or changes in behaviour across time periods.

Desireable features of entropy variable:

- Based on a large enough number of categories so that spend on many of them can reasonably be interpreted as chaotic (the 9 LBG tags seem insufficient for this, especially because most of them are vital life expenses). Use of auto tags or merchant seems preferable.
-