Spending profiles predict savings*

Fabian Gunzinger Neil Stewart

September 22, 2022

Contents

*We are grateful to Redzo Mujcic and and Zvi Safra for helpful comments. The research was supported by Economic and Social Research Council grant number ES/V004867/1. WBS ethics code: E-414-01-20. Gunzinger: Warwick Business School, fabian.gunzinger@warwick.ac.uk; Stewart: Warwick Business School, neil.stewart@wbs.ac.uk.

1 Methods

1.1 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]

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.

[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).

We thus do not exclude them.

1.2 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.

Table 1: Sample selection

	Users	User-months	Txns	Txns (m£)
Raw sample	271,856	7,948,520	662,112,975	124,573
Drop first and last month	265,760	7,406,482	643,851,490	121,098
At least 6 months of data	231,888	7,318,202	638,056,402	120,155
At least one savings account	144,309	4,788,799	445,037,084	90,541
At least one current account	141,514	4,723,402	439,698,170	89,629
At least £5,000 of annual income	54,248	1,636,887	172,146,281	33,921
At least 10 txns each month	48,302	1,447,121	155,751,277	30,559
At least £200 of monthly spend	42,274	1,249,750	139,156,150	27,471
Complete demographic information	34,657	1,072,446	119,534,253	23,142
Drop test users	34,581	1,067,566	118,984,718	23,018
Working age	33,884	1,043,727	116,923,733	22,251
Final sample	33,884	1,043,727	116,923,733	22,251

1.3 Dependent variables

Identifying savings transactions: 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 (similarly for debits).[3]

Dummy for savings txn in current month. Motivation: MPS (2018) finds that saving habit is often more important than amount saved.

1.4 Spending profiles

We define a user's spending profile as the distribution of the number of spending transactionas across different spend categories. To summarise these distributions, we calculate spending entropy, based on the formula proposed by Shannon (1948), who defines entropy as $H = -\sum p_i log(p_i)$, which sums, for all possible events, the product of the probability

[3]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.

of an event i occuring with the logarithm of that probability.[4] The base the logarithm is often chosen to be 2, though other choices are possible.[5] 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). In our context, we define the entropy of a user's spending profile in a particular period as:[6]

$$H = -\sum_{c \in \mathcal{C}} p_c log(p_c), \tag{1}$$

where $\mathcal{C}$ is the set of all spending categories, p_c the probability that an individual makes a purchase in spending category c, and log the base 2 logarithm. Higher entropy means that transactions are more equal across different spending categories, which makes it hard to predict the next transaction, whereas low entropy profiles have the bulk of transactions in a few dominant categories (such as groceries and transportation) and have relatively few transactions in other categories.[7] For simpler interpretation of our regression coefficients below, we standardise entropy scores to have a mean of 0 and a standard deviation of 1.

We calculate entropy based on three sets of spend categories. The first measure is based on 9 spending categories used by Muggleton et al. (2020).[8] The second measure is based on our own, more fine-grained, categorisation into 48 different categories.[9] The third measure is based on merchant names, as labelled by Money Dashboard.

We also calculate spending category probabilities in two different ways. To calculate what we call "unsmoothed" entropy scores, we calculate the p_cs in Equation 1 as simple frequentist probabilities

$$p_c = \frac{f_c}{F}, \tag{2}$$

where f_c is the number of transactions in spend category c (the frequency with which c occurs) and $F = \sum_{c \in \mathcal{C}} f_c$ the total number of spending transactions. To avoid taking the log of zero for categories with zero transactions, the sum in Equation 1 is taken over categories with positive transaction counts only.[10] To calculate "smoothed" entropy scores,

[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] The choice of the base for the logarithm varies by application and determines the units. Base 2 means that information is expressed in bits. The natural logarithm, another popular choice, expresses information in *nats*.

[6] We omit individual and time subscripts to keep notation simpler.

[7] For further discussion on how to interpret Equation 1, see Appendix **??**.

[8] The precise mapping from MDB transaction tags into these 9 categories is available on Github.

[9] The precise mapping from MDB transaction tags into these 48 categories is available on Github.

[10] This is automatically handled by the entropy implementation of Python's SciPy package, which is what we use to calculate entropy scores.

we apply additive smoothing to calculate propabilities as

$$p_c^s = \frac{f_c + 1}{F + |\mathcal{C}|}, \tag{3}$$

where the size of set $\mathcal{C}$, $|\mathcal{C}|$, is the number of unique spending categories. Hence, additive smoothing simply adds one to the numerator and the number of unqiue spending categories to the denominator of the unsmoothed probabilities. Because categories with a zero transaction count will have a numerator of 1, the sum in Equation 1 will be taken over all categories.

One way to think of entropy is as a function of a number of simple components, and we can rewrite Equation 1 in a way that makes this transparent. Let $\mathcal{C}^+ = \{c : f_c > 0\}$ be the set of all spending categories with positive frequency counts (i.e. with at least one transaction) and $\mathcal{C}^0 = \{c : f_c = 0\}$ the set of all spending categories with a zero frequency count, so that $\mathcal{C} = \mathcal{C}^+ \cup \mathcal{C}^0$. Then, using our definitions of unsmoothed and smoothed probabilities above, we can write unsmoothed entropy as

$$H = -\sum_{c \in \mathcal{C}^+} \left(\frac{f_c}{F}\right) log \left(\frac{f_c}{F}\right), \tag{4}$$

and smoothed entropy as:

$$H^s = -\sum_{c \in \mathcal{C}^+} \left(\frac{f_c + 1}{F + |\mathcal{C}|}\right) log \left(\frac{f_c + 1}{F + |\mathcal{C}|}\right) - |\mathcal{C}^0| \left(\frac{1}{F + |\mathcal{C}|}\right) log \left(\frac{1}{F + |\mathcal{C}|}\right), \tag{5}$$

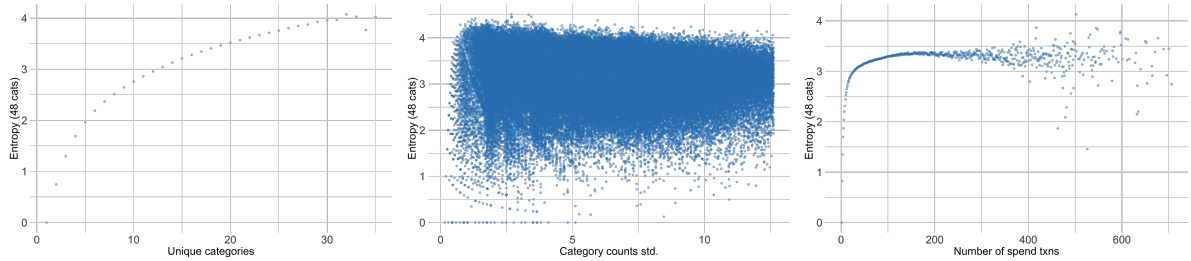
where the size of set $\mathcal{C}^0$, $|\mathcal{C}^0|$, is the number of all spending categories in which a user makes no transactions in a certain period. These expressions make clear that, by definition, unsmoothed entropy is a function of frequency counts of categories with positive counts only to avoid taking logs of 0, while smoothed entropy has two parts: the additively smoothed unsmoothed entropy, plus a constant term for each spending category with a zero frequency count.

The expressions also make transparent the three main components of both types of entropy: the number of spending categories with a non-zero frequency count ($|\mathcal{C}^+|$), the variation in the frequency counts of those categories (the variation of the f_cs), and the total number of transactions (F).[11]

Figure 1 shows the empirical relationship with our 48-categories-based unsmoothed

[11]The number of total spending categories, $|\mathcal{C}|$, also implicitly determines unsmoothed entropy (since it "scales" the number of categories with a positive frequency count, $|\mathcal{C}^+|$, as a given number of spending transactions are categorised into finer or coarser categories) and explicitly determines smoothed entropy, but it is exogenously given and does not depend on user behaviour. Also, the number of spending categories with zero counts $|\mathcal{C}^0|$ enters the expression for smoothed entropy, but we use it only in the hope that it makes the decomposition of entropy more transparent, when what it really is for an exogenously given $|\mathcal{C}|$ and any given $|\mathcal{C}^+|$ is $|\mathcal{C} \backslash \mathcal{C}^+|$.

Figure 1: Correlation of entropy with its components



Notes: Correlation of 48-categories-based unsmoothed entropy with its three main components: the number of unique spending categories with positive frequency counts (left), the standard deviation of those frequency counts (middle), and the number of total spend transactions (left).

entropy variable and these three components.[12] We can see that for the values we observe in the dataset, entropy increases monotonically in the number of unique spending categories with positive frequency counts, has no clear relationship with the standard deviation of those counts, and increases in the number of total spending transactions up to about 175 transaction, before being increasingly determined by other elements thereafter.

One slight limitation introduced by the imperfect transaction labelling in the MDB data is that entropy scores for high-entropy individuals will be biased downwards. This happens because unlabelled transactions tend to be transactions that are rare (i.e. not grocery or Amazon purchases), and it is high-entropy individuals that are more likely to engage in rare transactions. Because our analysis mainly relies on relative entropy levels, this is not of major consequence and we do not pursue this further.

1.5 Summary statistics

Table 2 provides summary statistics.

Table 2: Summary statistics

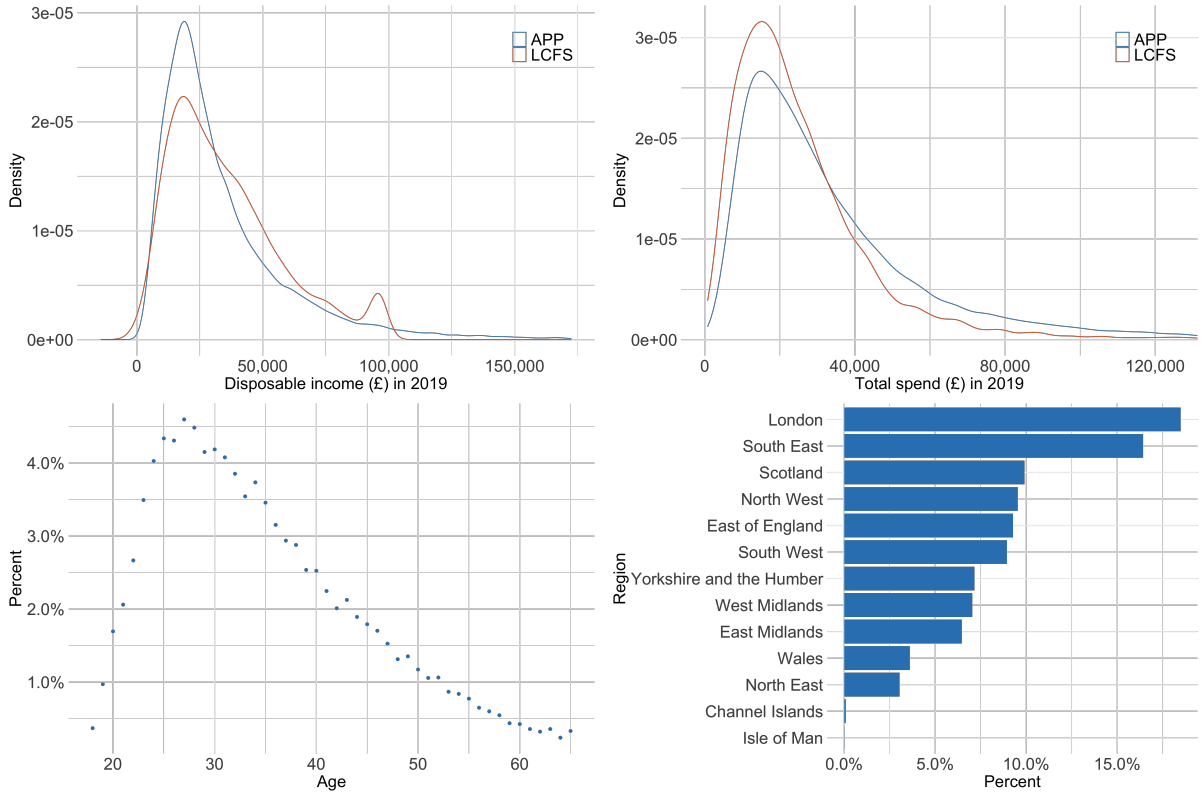
Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Year income	31.34	44.02	0.43	14.43	22.81	37.05	4,229.21
Income variability	1.23	4.87	0.00	0.21	0.54	1.16	559.67
Has income in month	0.98	0.15	0	1	1	1	1
Has savings	0.48	0.50	0	0	0	1	1
Month spend	2.70	2.65	0.20	1.16	1.91	3.19	17.44
Age	35.14	10.27	18	27	33	42	65
Female	0.39	0.49	0	0	0	1	1
Urban	0.85	0.35	0	1	1	1	1
Unique categories (9)	7.51	1.26	1	7	8	8	9
Unique categories (48)	15.15	4.43	1	12	15	18	35
Unique categories (Merchants)	22.99	9.77	0	16	22	29	85

Notes: Income and spend variables in '000s of Pounds, number of unique categories for spend transaction classification based on 9 categories, 48 categories, and merchant names.

[12] To highlight the main features of the relationships we have trimmed the component values at the 95th percentile.

Figure 2

Figure 2: Demographic characteristics of Money Dashboard users



Notes: The top left and top right panels show the distribution of disposable income and total spending in 2019, respectively, benchmarked against the 2018/19 wave of the ONS Living Cost and Food Survey (LCFS). The bottom left panel shows the distribution of age, the bottom right panel that of the regions.

Figure 3

1.6 Estimation

We estimate models of the form:

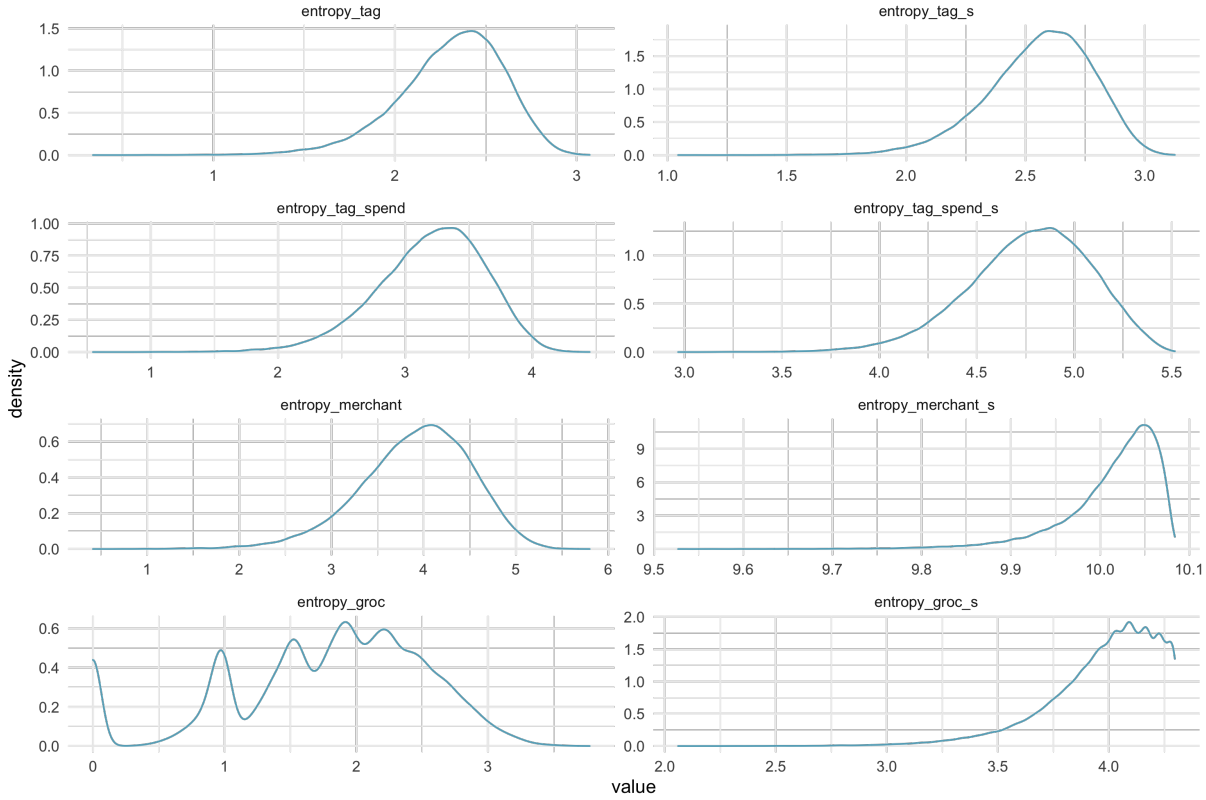
$$y_{i,t} = \alpha_i + \lambda_t + \beta H_{i,t} + x'_{i,t}\delta + \epsilon_{i,t}, \tag{6}$$

where $y_{i,t}$ is an indicator variable equal to one if individual i made one or more transfers to any of their savings account in year-month period t and zero otherwise, H_{it} is i's spending entropy in year-month period t, $x_{i,t}$ a vector of control variables, α_i an individual fixed effect, λ_t a year-month fixed effect, and $\epsilon_{i,t}$ the error term.

The vector of controls includes month spend, month income, an indicator for whether a user had positive income in a given month, and income variability, calculated as the standard deviation of month income over the previous 12 months.

Note that while we might in principle be worried about reverse causality, since making payments into savings accounts might lead to a non-zero count in an additional spend category and thus change entropy, this is not a concern here. As discussed in Section 1.3

Figure 3: Entropy distributions



Notes:

and Section 1.4, we define savings as inflows into savings accounts and define entropy based on the classification of spend transactions on current accounts. If a user pays money from their current into one of their savings account, such a transaction will usually be labelled in their current account as a transfer rather than a spending transaction, and thus not enter the calculation of their entropy score. In Appendix **??**, we provide robustness checks using lagged entropy scores, which produces very similar results.

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.

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: