BBVA Research

CLARITY AI

Creating Opportunities

N.º 21/05

Working Paper

Cash Vs Card Consumption Patterns in Mexico: A Machine Learning Approach

Jaime Oliver, Álvaro Ortiz, Tomasa Rodrigo,
Saidé Salazar and Ignacio Tamarit

May 2021

Cash Vs Card Consumption Patterns in Mexico: A Machine Learning Approach

Jaime Oliver (Clarity AI), Alvaro Ortiz (BBVA Research), Tomasa Rodrigo (BBVA Research), Saidé Salazar (BBVA Research), Ignacio Tamarit (Clarity AI)[1]

May 2021

Abstract

The analysis of household consumption patterns is relevant for social welfare, policy design, and economic analysis. Traditional empirical analyses of consumer behavior are based on household consumer surveys, which provide an incomplete picture and are particularly flawed when high levels of informality are present. This paper proposes a novel methodology combining high frequency card transaction data and point-of-sale (POS) data from cash operations registered at convenience stores. We study the changes in consumption patterns in Mexico relative to variations in income, including changes in the items consumed and the payment channel (cash vs. electronic payment). In particular, we analyze how individuals allocate their card and cash purchases using a variety of econometric and Machine Learning Models. We leverage recent developments in model explainability based on Shapley values to deep dive into the Random Forest champion model, which achieves R^2 scores above 0.92. The results show that the most relevant variables to increase the expenditure by card relative to cash are the changes in income, living in an urban center, and the financial deepening effects. Whereas income produces critical nonlinear effects for both types of transactions, urban levels and financial deepening have an increasing influence in card operations.

Keywords: consumption patterns, cash, electronic payments, big data, consumption elasticity, Machine Learning, Shapley Values

JEL classification: C32; D12; O17; O54.

1: We really thank Marta Rivera Alba and Tatiana Dávila for their contribution to the paper and assistance, without them this paper would not have been possible.

1. Introduction

The analysis of individuals and household consumption patterns can become an important tool to improve the lives of those in need. Through a detailed geographical and time data analysis, resource shortages and surpluses can be identified. In turn, this information can also guide policy makers to tackle public and private initiatives to achieve the UN SDG goals of zero hunger and no poverty. Further disaggregation of consumption profiles by region or gender could inform other SDG goals like gender and social equality and economic growth or education.

Consumption patterns are traditionally analyzed via household surveys. While this approach has many advantages, it also has some shortcomings; it is very time and resource consuming, which leads to a low frequency update that may fail to capture critical events.

In this paper, we propose the use of real time and high definition data to track consumption. We combine two datasets: a large dataset of card payments from the Big Data of a Bank and a dataset from an integrated Point of Sale (POS) platform that allows small businesses to track their transactions--most of them in cash. We use the former to proxy digital transactions and the latter to proxy cash consumption. This novel approach allows us to analyze two complementary views of the economy, having a comprehensive understanding of consumption patterns from different payment channels. Besides, the high granularity of the data, both in terms of time and transaction resolution, provides valuable insights for policy making.

Consumption patterns reveal people's preferences among the different categories of products and services. Using econometric and machine learning models, we can classify how the different socioeconomic groups perceive each consumption category as a basic or superior goods and estimate whether these groups are over or underspending in a given category. This analysis will allow us to identify the relevant consumption categories to focus in order to achieve a larger social impact. Potentially, the temporal dimension of this proxy consumption data could enable the analysis of changes in consumption both at a category and brand level, in real time, due to changes in a consumer's available budget, or personal or social events.

Our specific case study is Mexico, where our combined methodology can add significant value due to the importance of the informal economy as well as its high inequality levels (see Annex 1). The financial deepening in Mexico is relatively low, with most of its low-income population remaining unbanked, and a substantial part of their transactions being made through cash transactions. Low income levels are commonly associated with lower education levels, fewer worker skills, smaller firm size (familiar) and temporary jobs. These structural characteristics of the economy favor informality and constitute an important factor determining the use of payment channels. As a consequence, from a population of 130 million, only 38.7% of Mexican individuals have an account with a financial institution. To address this issue, the Mexican government implemented the CoDi program[2] to promote the use of electronic payments. High resolution data enables the impact of these types of policies to be analyzed and the heterogeneity of consumption patterns across regions, sectors and client's features to be studied.

The paper is organized as follows, section 2 describes the sources data for cash and card transactions used in the project as well as the main explanatory variables representing income level, urbanization and financial deepening. The section also describes an exploratory data analysis.

2: Program implemented by the Mexican Central Bank to promote the use of electronic payments to replace cash transactions in the economy, introducing a cell phone payment platform that uses QR codes. More information in the Annex.

Section 3 shows the results of a linear model for both card and cash transactions as a function of income, urban, geographical and financial deepening levels as control variables. Once the model is estimated, we also estimate the semi-elasticity of consumption to income for different goods and means of payment.

In section 4, we explore the ability of Machine Learning models to capture the non-linearities of consumption patterns. We explore some alternative models as Linear Regression, Logit, Random Forest and SVM and we translate the relevance of explanatory variables using Shapley values.

The last section is dedicated to comment on the main results.

2. Data

2.1 Data description

To characterize the Mexican population's consumption patterns, we combine information from purchases made in cash at convenience stores with card spending using BBVA cards. The cash data were recorded using a simple app from Frogtek, a for-profit social venture dedicated to creating business tools for small shopkeepers in emerging markets. Frogtek develops its own mobile applications for small shops and customers at the bottom of the value chain. It has also created a cloud platform with services for partners like payment providers, market researchers and consumer packaged goods companies. One of their flagship projects is Tiendatek, an integrated POS platform that allows small businesses to track their sales and inventory, register transactions, obtain useful metrics and charge credit cards.

The Frogtek sample includes 62 million transactions from 2016 to 2018, from 1835 shops in 12 states in Mexico and about 26,000 different products. We aggregated the data to cover 13 COICOP consumption categories (including food, alcoholic beverages, transport, equipment and health) while the original data set identifies 171 types of products. We were able to attribute an income level to the cash consumers by using the average income of the municipality. Furthermore, we assigned consumption to individual consumers by aggregating consumption frequency data by category obtained from the National Institute of Statistics & Geography in Mexico (INEGI).

The data from card spending is obtained from BBVA transaction data. BBVA is the largest banking institution in Mexico with a 20% market share, and with presence in seven other countries[3]. We analyze all transactions registered from a BBVA credit or debit card, as well as BBVA POSs, from 2017 to 2018. This includes 3.2 billion transactions, disaggregated across 32 states and 1987 municipalities and 344 types of consumption categories. Moreover, we use the clients' socioeconomic features to separate the data into five different income levels.

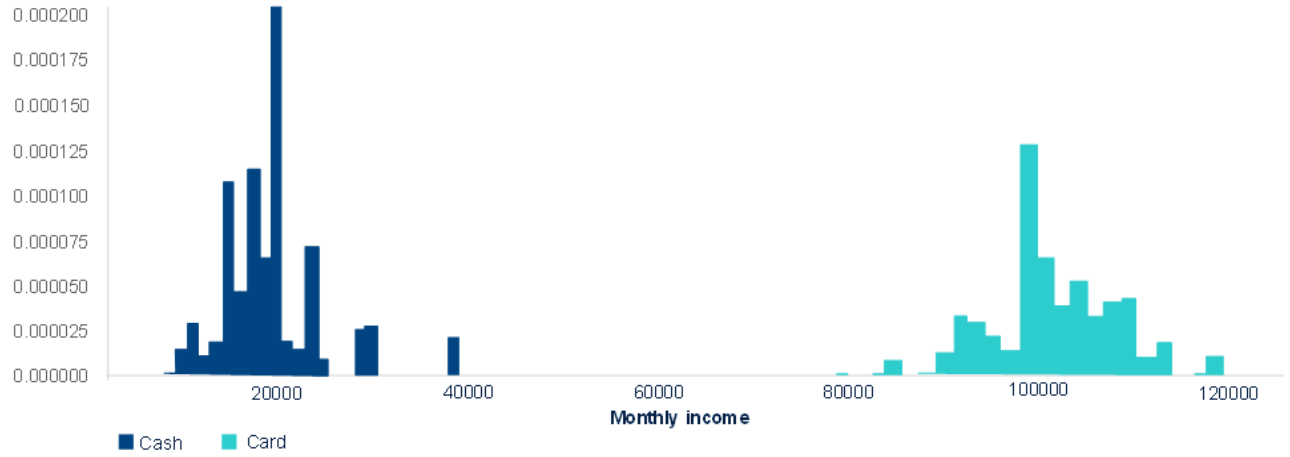
3: BBVA is based in Spain, Turkey, Mexico, Peru, Colombia, Argentina and Venezuela.

2.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) reveals significant differences in the transaction made by cash and card at different income levels (see Figure 1). The card transactions data are more frequent at higher income levels while cash transactions are more normal for lower income individuals. Hence, combining both data sources is key to capturing complementary socioeconomic sectors.

The granularity of the data enables us to integrate of POS and Bank data. We aggregate the data at monthly levels and focus on 43 municipalities for which we have data for both card and cash consumption. First we check for the existence of biases. Figure 2 compares the distribution of our sample population with the Mexican one. We observe no population bias in terms of the degree of urbanization of the municipalities covered, as well as between cash and card data. For this comparison, we take into account five categories of municipalities[4] based on the number of inhabitants: rural (less than 5,000 inhabitants), transitioning (from 5,001 to 15,000 inhabitants), semi-urban (from 15,001 to 50,000 inhabitants), urban (from 50,001 to 300,000 inhabitants), semi-metropolitan (from 300,001 to 1 million inhabitants), and metropolitan (more than 1 million inhabitants).

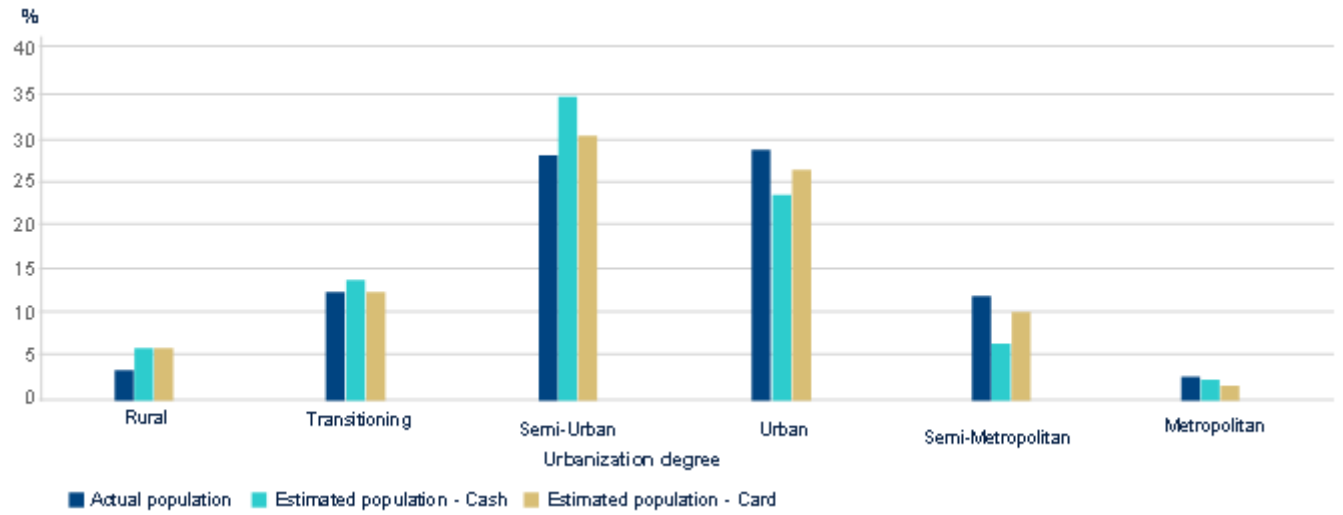
Figure 1. **INCOME DISTRIBUTION OF CASH/CARD DATA**



We observe that card data covers high income municipalities, while cash data has a better coverage of low income municipalities. Thus, the complementary nature of the two datasets allows us to cover different socio-economic segments of the population.
Source: BBVA and Frogtek

4: All these categories correspond to the classification of municipalities by level of urbanization used by the National Banking and Securities Commission in Mexico (CNBV) in its annual financial inclusion reports.

Figure 2. **DEGREE OF URBANIZATION DISTRIBUTION OF CASH/CARD DATA**



Distribution of actual and estimated populations across degree of urbanization. We observe that both the cash and card datasets are representative of the actual population, with a slight trend of cash data towards rural areas and card data towards metropolitan areas.
Source: *CNBV,* BBVA and Frogtek

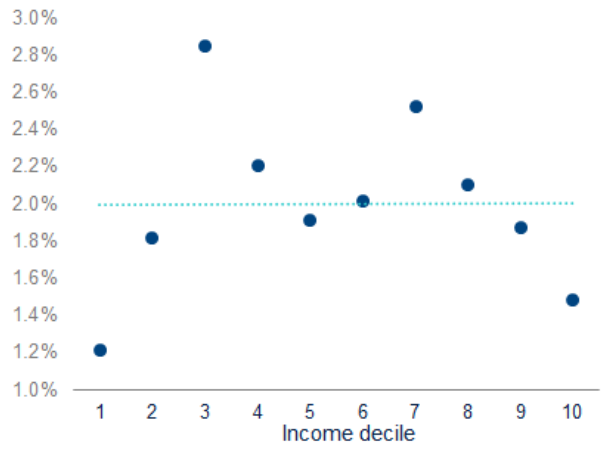
We use the cash and card data to build an indicator of each individual's budget share, that is, the cash and card expenditure by category per person divided by the income level. In the case of the cash data. In line with official INEGI reports, we assume that each person makes six purchases per month in a shop that uses the Frogtek app and we assume that the average income corresponds to his/her municipality (data also from INEGI). For card data, we consider the total amount consumed divided by the number of BBVA clients in each category and we use the income level of each client to calculate their budget share.

The evolution of card and cash data trends is closely related to the evolution of informality in Mexico. The informal economy made up 26.2% of GDP in 2019 (see Figure A2. in Annex 1) and around 56% of the working population. At the same time, the amount of banknotes and coins held by the public reached 62.7% of GDP, and nearly 95% of purchases of 500 Mexican pesos or less (USD 25 or less) were made in cash. By geography, we notice that the states with the lowest incomes have the highest rates of informality in the country (see Annex 1).

The individual and households' decisions to consume with card or cash, measured as percentage of the total budget, depends on multiple factors. One important factor looks to be the Income level. The analysis of the budget share by income deciles for card and cash spending show some interesting patterns (Figures 3 and 4).

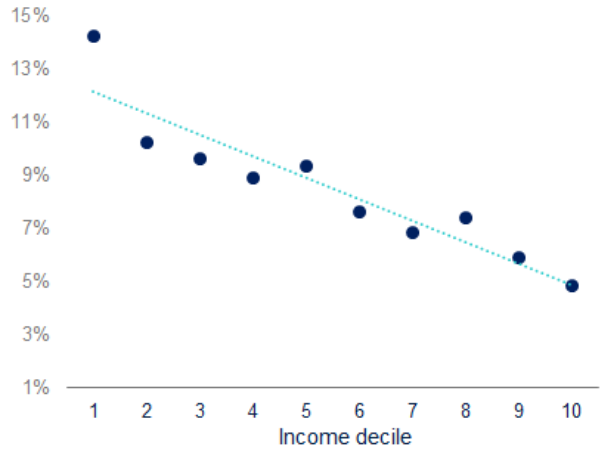
While card spending is less sensible to income decile, we find a significant negative relationship between cash spending and income, where the highest income decile has a cash spending budget share almost 10 percentage points lower than the lowest income decile (Figure 4). In the case of card spending, however, the results show a higher budget share among the mid deciles, but similar shares are observed among the lowest and the highest ones (Figure 3).

Figure 3. **CARD SPENDING BUDGET SHARE BY INCOME DECILE**



Source: own calculations. BBVA and Clarity

Figure 4. **CASH SPENDING BUDGET SHARE BY INCOME DECILE**



Source: own calculations. BBVA and Clarity

2.3 Regional analysis

Taking advantage of the granular data we have by municipality, we analyze the card and cash evolution regionally and we find a heterogeneous pattern across regions and time for the 43 analyzed municipalities. First of all, within the same state, high income municipalities are usually associated with a higher increase of card spending over time, and low income and densely populated municipalities with a higher increase of cash spending. These findings are in line with the well documented negative correlation between income level and informality at a regional level (see Annex 1). The differences across municipalities in the same state underlines the poor income distribution that also characterizes the country (Figures 5 and 6).

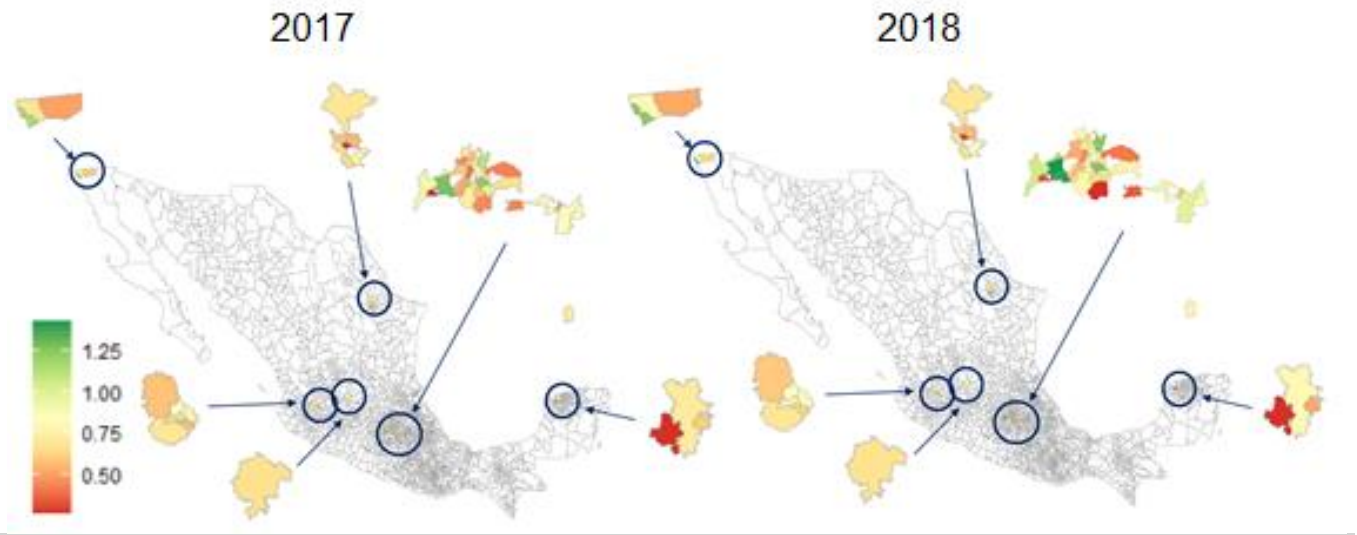
We found heterogeneous patterns across the different regions of Mexico City: the municipalities with the lowest income per capita show a higher increase in cash spending (Tlahuac, Tlalpan, Gustavo A. Madero), while the wealthiest register the lowest growth in cash usage (as seen in Miguel Hidalgo), or even a decrease (Iztacalco). Overall, most of the municipalities (except Iztacalco and Milpa Alta) show an increase in cash budget share during the period (2016–2018). On the contrary, the card budget share showed almost zero growth in all Mexico City's municipalities covered by this analysis during 2017–2018. Overall, in Mexico City 44 of each 100 workers belong to the informal sector.

In Nuevo León, the third wealthiest state of the country, the municipality with the highest income covered by this study (San Nicolás de los Garza) showed the lowest cash budget share of the state sample in 2018, while one of the poorer municipalities (Juarez) shows the highest cash usage. Overall, in Nuevo Leon 35 of each 100 workers belong to the informal sector. As reference, the average rate of informality for the main cities in the country is 41%, 13 pp below the national figure (54%), which suggests that informality is concentrated in less urbanized areas (see Annex 1 for more details).

In Yucatán, the state with the lowest income from our sample, the municipality of Merida (the state's capital city) shows a higher increase in card spending during the period analyzed compared to the municipalities of Kanasin and Uman, which also registered positive variations in card budget share but smaller. Interestingly, Uman is a

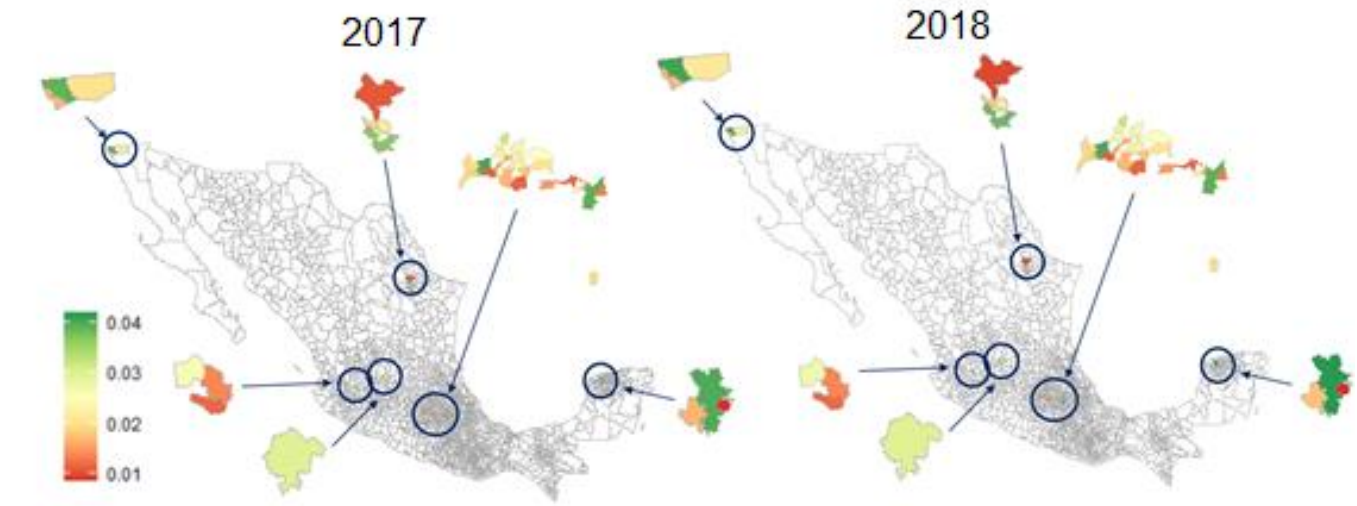
unique case, being the state with one of the lowest use of cash in the sample, just after San Nicolas de los Garza in Nuevo Leon.

Figure 5. **CASH SPENDING BUDGET SHARE BY MUNICIPALITY**



Source: own calculations. BBVA and Clarity

Figure 6. **CARD SPENDING BUDGET SHARE BY MUNICIPALITY**



Source: own calculations. BBVA and Clarity

3. A Linear Model of Mexican Consumption patterns: Card vs Cash

After the exploratory data analysis section, we train statistical and machine learning models to explain consumption patterns in Mexico. This will provide us with complementary insights about the consumption patterns and payment transactions in Mexico and quantify the relevance of the variables.

3.1 A linear Model of Mexican consumption patterns

To analyse the relationship between income versus card and cash transactions, we model the share of card or cash transactions relative to income $\omega_{it} = \frac{C_{it}}{I_{it}}$ as a function of alternative exogenous variables in the following panel data specification

$$\omega_{it} = c + \beta_1 log\,(I_{it}) + \beta_{iFD} X_{it}^{FD} + \beta_{iU} X_{it}^{U} + \beta_{iG} X_{it}^{G} + u_{it} \qquad (1)$$

where *t* is the time index, *i* represent municipalities computed by averaging individuals, C_{it} is consumption done in cash or card, and I_{it} is the income level. The model includes an intercept, fixed effects, and a set of explanatory variables: the income Level (I_{it}) effect and a set of control variables representative of financial deepening (X_{it}^{FD}), level of urbanization (X_{it}^{U}),and geography (X_{it}^{G}).

The income level is introduced in logs $log\,(I_{it})$, and as the dependent variable represent a share of the total income ω_{it}, the coefficient β_1 stands for the semi-elasticity of consumption to income. As a proxy for financial deepening (X_{it}^{FD}), we consider two variables: deposits contracts per 10,000 inhabitants, and credit contracts per 10,000 inhabitants. Both variables are published by the CNBV and monitor the evolution of financial services use or intensity per municipality. The degree of urbanization (X_{it}^{U}) is also an important control variable as there are reasons from the demand side (level of income) and the supply one (banking development), which can affect the results. For this we introduce categorical dummies representing a metropolitan or rural area. Finally, we also add geographical variables (X_{it}^{G}) to control for differences in economic development across municipalities associated with location.

3.2 Results

The results for the model estimates show some important differences between the regressions for card and cash spending. The semi-elasticity of income shows a positive and significant effect of income level for card spending reflecting that as the income rises, the share of goods consumed by card also increases (Table1). The opposite happens in the cash regressions showing that cash expenditure decreases with the level of income. The negative effect in cash consumption is slightly higher and more significant than the one in the card regressions. This confirms in quantitative terms the relationships observed in the data exploratory analysis where the slope for the cash is clearly negative and significant while the data for card transactions is less clear and somehow flatter.

Most of the control variables are all significant too but with different signs in the card and cash regressions. The level of urbanization[5] (these variables are included with respect to the rural area) has a positive effect for card spending, while it is negative for cash spending. The asymmetry is also present in the regional dummies of the northern and southern regions of the country relative to central regions. There is a positive relationship of North and South relative to center in the card regressions while the effect in the case of cash is negative and significant. Finally, financial inclusion has a positive effect in both cash and card spending, although it's the least significant of all the explanatory variables (Tables 1 and 2).

Table 1. **ESTIMATION RESULTS USING LINEAR MODEL FOR CARD SPENDING**

	Parameter	Std. Err.	T-stat	P-value
Intercept	-0.0731	0.0302	-2.4253	0.0154*
log(Monthly income)	0.0071	0.0027	2.5725	0.0102*
log(Deposit contracts per 10000)	0.0021	0.0005	4.1337	0.0000***
log(Credit contracts per 10000)	-0.0020	0.0006	-3.2251	0.0013***
Metropolitan area	0.0178	0.0017	10.184	0.0000***
Urban area	0.0085	0.0017	5.0152	0.0000***
Geography North	0.0052	0.0006	8.7008	0.0000***
Geography South	0.0041	0.0007	5.9053	0.0000***

Regression results of the linear panel data model for card budget share (R^2 = 0.39). We observe a positive and significant income elasticity of demand. Control variables—financial inclusion, level of urbanization and geography—also have significant effects.
Source: own calculations. BBVA and Clarity

Table 2. **ESTIMATION RESULTS USING LINEAR MODEL FOR CASH SPENDING**

	Parameter	Std. Err.	T-stat	P-value
Intercept	0.9288	0.0223	41.625	0.0000***
log(Monthly income)	-0.0907	0.0028	-32.843	0.0000***
log(Deposit contracts per 10000)	0.0024	0.0014	1.7246	0.0848
log(Credit contracts per 10000)	0.0043	0.0018	2.3557	0.0186*
Metropolitan area	-0.0056	0.0053	-1.0628	0.2880
Urban area	-0.0172	0.0051	-3.3542	0.0008***
Geography North	-0.0017	0.0017	-1.0299	0.3032
Geography South	-0.0125	0.0020	-6.3674	0.0000***

Regression results of the linear panel data model for cash budget share (R^2 = 0.53). We observe a negative and significant income elasticity of demand. Control variables—financial inclusion, level of urbanization and geography—also have significant effects.
Source: own calculations. BBVA and Clarity

5: The level of urbanization corresponds to rural, in transition, semi-urban, urban, semi-metropolis and metropolis (source: National Banking and Stock Commission).

3.3 Classifying Normal and Superior goods through Income semi-elasticities

Economic theory considers normal goods as the products and services that are purchased by consumers regardless of their income level. Their elasticity to income is positive but lower than superior goods, so that the share of expenditure dedicated to these goods decreases as income rises--their semi-elasticity of demand is negative[6]. On the contrary, superior goods have a positive and higher elasticity to income than normal goods as their consumption rises more than the latter as income increases. Contrary to normal goods, their semi-elasticity to income is positive as the share of these goods increases when income rises.

The estimation of the semi income elasticity of demand allows these normal and superior goods to be identified. Thanks to the high granularity of data, we can estimate the model by product category level and identify normal and superior goods depending on the value of the coefficient representing the semi-elasticity of demand of the different goods. We find that most goods classified as normal (negative coefficient for semi-elasticity) were paid with cash (Table 3). We found Food, Alcoholic Beverages and Tobacco among them. On the other hand, most goods classified as superior (positive coefficient for semi-elasticity) were purchased with card--Travel, Clothing, Restaurants and academia presenting the highest values.

Table 3. **SEMI ELASTICITY OF DEMAND OF NORMAL AND SUPERIOR GOODS**

Superior Goods

Consumption category	Source	Income elasticity	p-value
Others	Card	0.4968	0.0000E+00
Travel	Card	0.2675	1.7292E-09
Clothing	Card	0.1767	0.0000E+00
Restaurants	Card	0.1505	0.0000E+00
Academia	Card	0.1011	2.7325E-08
Other services	Cash	0.0692	6.1209E-05
Pets	Card	0.0285	5.5511E-15
Office	Card	0.0275	2.3891E-04
Shoes	Card	0.0206	1.9322E-03
Beauty	Card	0.0206	6.6613E-16
Music	Card	0.0139	2.9060E-04
Jewelry	Card	0.0087	1.4360E-09
Cash withdrawals	Card	0.0037	1.7764E-15
Books	Card	0.0033	6.3647E-04

Normal Goods

Consumption category	Source	Income elasticity	p-value
Home	Card	-0.0182	4.7010E-02
Material Construction	Card	-0.0591	6.6576E-08
Clothing	Cash	-0.1139	4.9960E-13
Transport equipment	Cash	-0.1757	0.0000E+00
Maintenance and repair of the dwelling	Cash	-0.5071	0.0000E+00
Food	Cash	-0.622	0.0000E+00
Personal care and effects	Cash	-0.6622	0.0000E+00
Non-alcoholic beverages	Cash	-0.6646	0.0000E+00
House, garden and pets	Cash	-0.6709	0.0000E+00
Food	Card	-0.7721	0.0000E+00
Alcoholic beverages	Cash	-1.5388	0.0000E+00
Tobacco	Cash	-1.8128	0.0000E+00

Consumption categories classified as superior goods with a positive income semi elasticity of demand. These categories account for 37% of all the categories, and the vast majority are paid by card. Source: own calculations. Consumption categories classified as normal goods with a negative income semi elasticity of demand. These categories account for 32% of all the categories, and the vast majority are paid with cash.
Source: BBVA and Clarity

6: Deaton, A.; Muellbauer, J.; others. Economics and consumer behavior; Cambridge university press: New York, NY, 1980.

4. Moving beyond linearity to test non-linear relationship in the data

4.1 Estimation of Machine Learning Models

Linear models have the advantage of interpretability. However, the relationship between budget share, income and the rest of control variables, could be non-linear and more complex. To test this hypothesis we fit a suite of models including Linear Regression (OLS), Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines, as well as a Logit model for which we binarize card and cash expenditure by one if the share is higher than the median share and zero otherwise.

We train the models using the 80% of data and select the best one based on five-fold cross-validation errors using a standard hyper parameter tuning schema. We use the remaining 20% of the data to test the model and estimate the generalization error. The best selected model from the training set is used to predict the unseen budget shares in the test set. We do this for each of the cash and card datasets and record the accuracy in the case of the logit model (the ratio of correct forecasts relative to the total of the alternative models) and performance (R^2) of the models in their respective test samples for the machine learning models.

The results in table 4 show the performance of the different models. The first important result is all of the models including Logit, Decision Tree, Gradient Boost and Random Forest present better results than the linear benchmark. This is already indicative that non-linear specifications can improve our linear model. A traditional specification in terms of the Logit Model (0.78 card and 0.82 for cash) has a good accuracy too.

Table 4. **MODELS ACCURACY / PERFORMANCE SCORE FOR CARD AND CASH SPENDING**

	Random Forest	Gradient Boosting	Decision Tree	Logit	OLS
Card Spending	0.93	0.83	0.70	0.78	0.39
Cash Spending	0.92	0.78	0.75	0.82	0.53

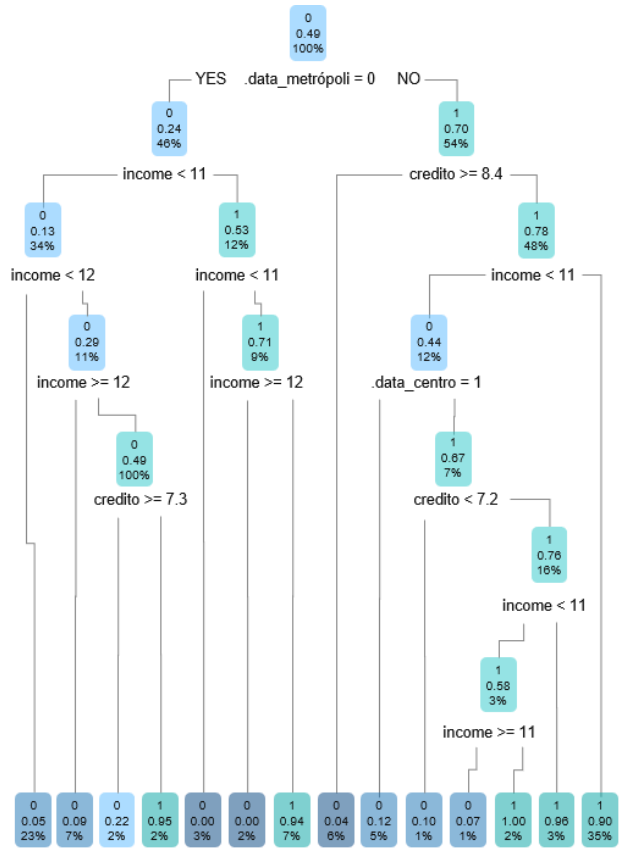
Source: own calculations. BBVA and Clarity

The best regressor results a Random Forest, a non-linear model based on the aggregation of multiple decision trees[7]. Remarkably, its performance is about 100% better than the linear benchmarks, obtaining R^2 of 0.93 and 0.92 for the Card and Cash models respectively. This undoubtedly points out to the existence of meaningful non-linearities.[8]

7: Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics
8: The Random Forest model copes with different non-linear relationships, which are difficult to analyze using traditional approaches. It generates a large number of regression trees, each of them calibrated on a bootstrap sample of the data. Each node is split using a subset of randomly selected predictors. For predicting the value of a new data point, the data is run through each of the trees in the forest and each tree provides a value. The model prediction is then calculated as the average value over the predictions of all the trees in the forest.

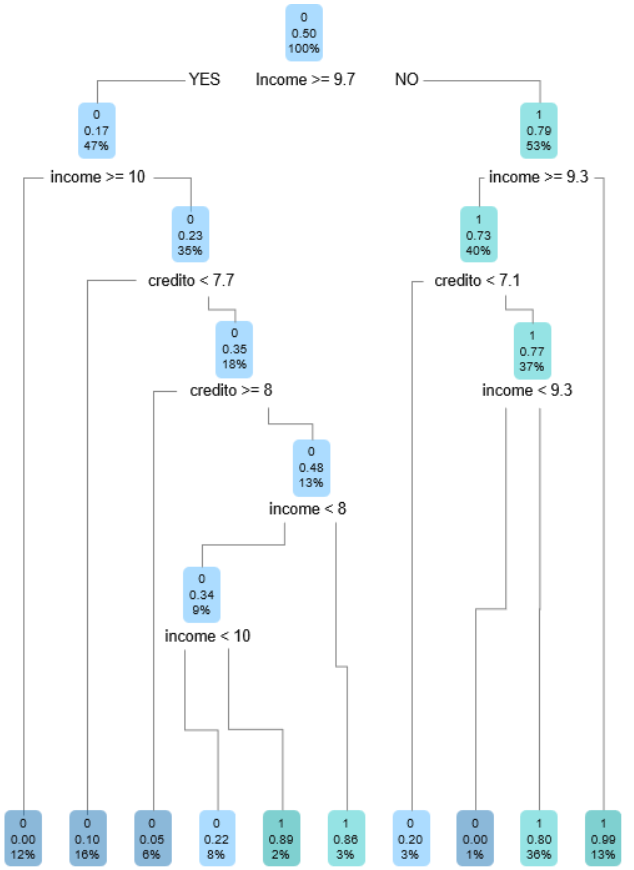
To gain insights on the type of non-linearities present in the data we turn now to analyze a single Decision Tree. Notice that even though its performance is lower than the Random Forest (as usually is), it is still much higher than the benchmarks[9].The card spending tree (Figure 7) reveals that living in a metropolitan area, financial inclusion and a high level of income are the most important variables. As observed, the model is able to capture the non-linearities at the higher levels of financial deepening (the right branch of the tree) between credit and income (lower levels of the tree). In the case of cash expenditure (Figure 8) the most discriminatory variable is income as the decision tree shows. In this case the non-linear relationships are at lower income levels (left branches of the tree) when interacting with financial deepening.

Figure 7. **CARD SPENDING DECISION TREE**



Source: own calculations. BBVA and Clarity

Figure 8. **CASH SPENDING DECISION TREE**



Source: own calculations. BBVA and Clarity

9: We should take into account that these decision trees are just one example to illustrate variables' relationships, but final results from the random forest approach can vary given it considers multiple combinations of decision trees. However, we find that, even though these trees have only depth 6, they account for 80% (0.75 R2) and 90% (0.86 R2) of the variability explained by the best card and cash models respectively, so the interpretation insights derived from this example are closed to the final output.

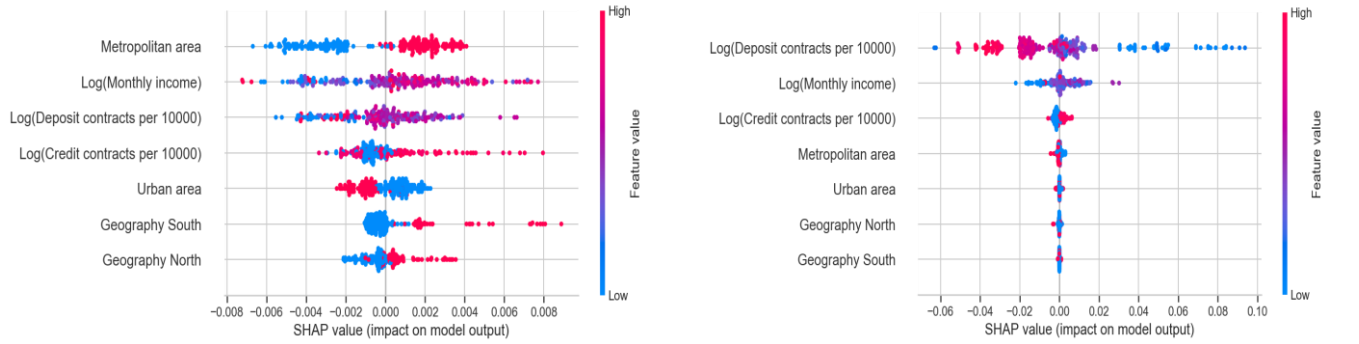
4.2 Identifying Machine Learning Values through Shapley values

While the Machine Learning Algorithms are better suited to capture the non-linear relationship between the variables, they also have some limitations. Notably, it might be hard to quantify the relationship between the variables --- a chief reason why they have often been referred to as “black box” models.

To gain some explicability on these models, SHAP values[10] (inspired by the game theory concept of Shapley values) assign the fair contribution of the explanatory variables to a prediction by averaging its marginal contributions. Importantly, SHAP values capture the drivers that push up or down a given estimate with respect to a baseline. Figures 9 show the SHAP value (x-axis) for any of the explanatory variables (y-axis) the x-axis for card and cash spendings.

The variables are ordered in terms of relevance and the color represents how high (red) and low (blue) values of the variable impact the model. In the case of card transactions (left graph) the higher values for income (in red) are associated with negative values of the dependent variable (the expenditure share) in line with a negative semi elasticity, while the opposite happens in the card spending model where the semi-elasticity is positive but less clear as we find a mild and positive effect for income. In a nutshell, we see that there is strong evidence of a nonlinear relationship between income and budget share captured using machine learning techniques. These effects are directionally consistent with their linear counterparts, although it was milder using the linear approach.

Figure 9. **SHAP SUMMARY PLOT USING NON-LINEAR MODEL FOR CARD & CASH SPENDING**



Source: own calculations. BBVA and Clarity

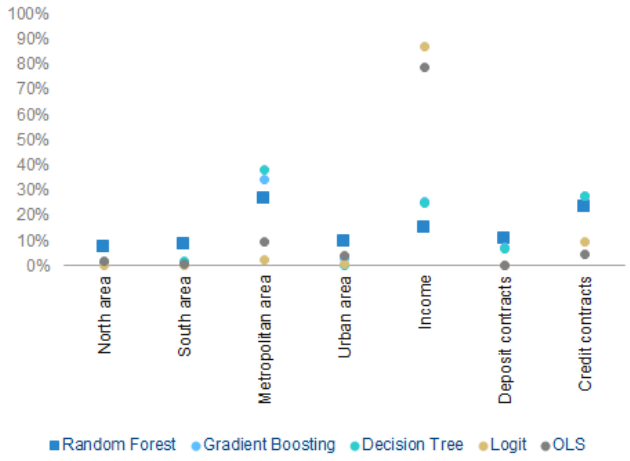
One important property of the SHAP values is that they are additive, in the sense that the independent contributions of the variables are equal to the total variation of the dependent variable. Figures 10 and 11 show the contribution of the variables for the alternative models tested in this exercise showing an important difference between the linear and non-linear models (data table can be found in the annex 2, Table A1).

Particularly, for the case of card spending, while linear and logit models show the relevance of the income variable, the non linear algorithms capture the influence of living in metropolitan areas and financial deepening representative variables. In the case of cash spending, the differences between the linear approximation and machine learning models are significant too (Figures 12 and 13). However, the relevance of Income is maintained in both linear approximations and the machine learning alternatives (although lower) while financial inclusion (credit

10: Lundberg, Scott, and Su-In Lee. "A unified approach to interpreting model predictions." arXiv preprint arXiv:1705.07874 (2017).

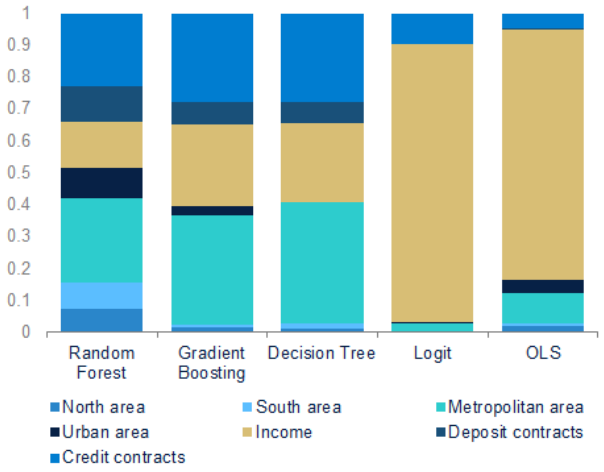
contracts) and, to a lesser extent, geography are still relevant but lower than in the cash transactions case (data table can be found in the annex 2, Table A2).

Figure 10. **SHAP SHARE COEFFICIENTS FOR CARD SPENDING MODELS BY VARIABLE**



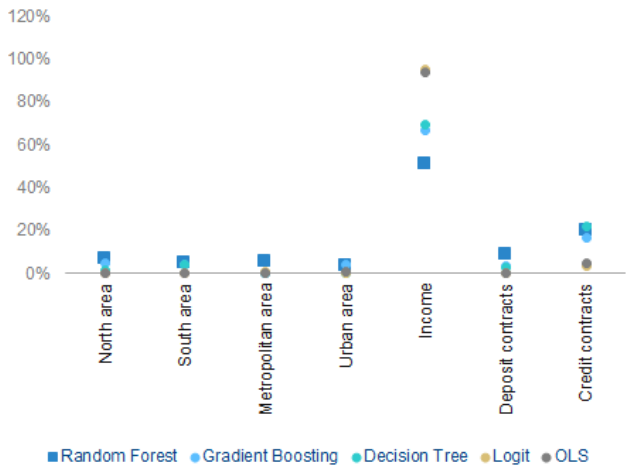
Models' order is based on models performance, from the best performance (random forest) to the worst one (OLS).
Source: own calculations. BBVA and Clarity

Figure 11. **SHAP SHARE COEFFICIENTS FOR CARD SPENDING MODELS BY MODEL**



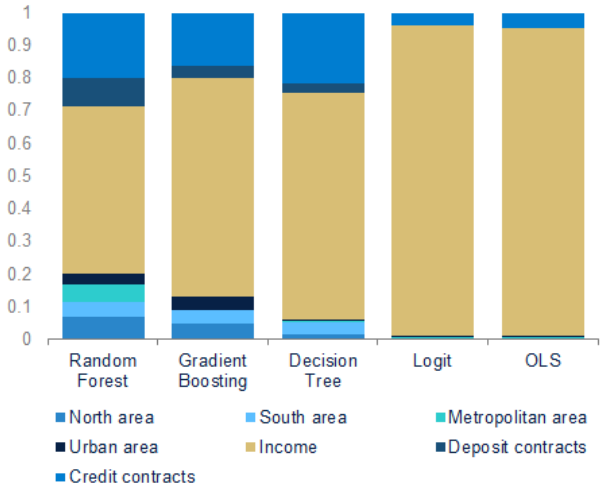
BBVA and Clarity. Models' order is based on models performance, from the best performance (random forest) to the worst one (OLS).
Source: own calculations. BBVA and Clarity

Figure 12. **SHAP SHARE COEFFICIENTS FOR CASH SPENDING MODELS BY VARIABLE**



BBVA and Clarity. Models' order is based on models performance, from the best performance (random forest) to the worst one (OLS).
Source: own calculations. BBVA and Clarity

Figure 13. **SHAP SHARE COEFFICIENTS FOR CASH SPENDING MODELS BY MODEL**



BBVA and Clarity. Models' order is based on models performance, from the best performance (random forest) to the worst one (OLS).
Source: own calculations. BBVA and Clarity

Conclusions

Consumption data is crucial for social welfare, policy making and economic analyses. However, the traditional survey data has some well-known shortcomings. In this work we propose an alternative approach using Big Data to track consumption combining two complementary datasets, one registering cash operations and another one registering card transactions and electronic payments. The cash consumption dataset is derived from digitized POS receipts, while the one used for card consumption comes from transactional data. Both datasets provide valuable insights for the analysis thanks to their high time, geographical, and product-level granularity.

Our data analysis points to the existence of significant differences among municipalities within the same state, linked to income level and the size of the informal economy in the region. Poorer and densely populated municipalities report the highest cash usage and/or the highest increase in cash budget share. Wealthier municipalities, on the other hand, register less cash spending and higher card spending.

The combined use of data from card and cash transactions has proven to be an effective tool for characterizing consumption needs. Our analyses reveal that income has significant linear and nonlinear effects on card and cash spending. The higher the income, the higher the level of bancarization and the lower the cash spending—even when controlling for financial inclusion, level of urbanization, and geography. In addition, we were able to identify which consumption categories are perceived as necessity or luxury goods, and the preferred payment methods for each of them. Non-surprisingly, most of the necessity goods detected are paid with cash. This goes along with the fact that convenience stores are the ones selling these types of goods in our sample. This technique to discriminate between the two types of consumption can be of great value, for example, in creating policies to protect specific sectors.

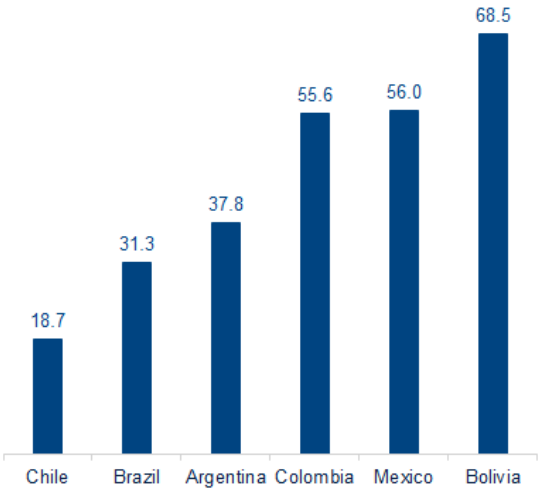
This paper also shows the importance of using Big Data and data science techniques to capture non-linear relationships between variables, getting complementary insights to more traditional approaches. The analysis of card and cash data shows different consumption patterns for the Mexican population depending on income, urban and rural and financial deepening levels. However, Machine Learning models show that patterns of consumption could be different depending on the payment transactions. Particularly, the influence of income looks to be more important in the case of cash transactions while the urbanization and financial deepening effects are well captured in the Machine Learning Models.

Annex 1 - Informality

In recent years the Mexican Central Bank has promoted public-policy initiatives aimed at encouraging the use of digital money or electronic payments to replace cash transactions in the economy. One such program is CoDi, a cell phone payment platform that uses QR codes. The initial phase of CoDi started on September 30, 2019, which included the technological adoption of the platform by the commercial banks and dissemination of information among the population for users enrolling in it (Banxico 2020). By December 2020, there were 7.4 million accounts validated through the program, 64% of which were validated by BBVA (CoDi 2021).

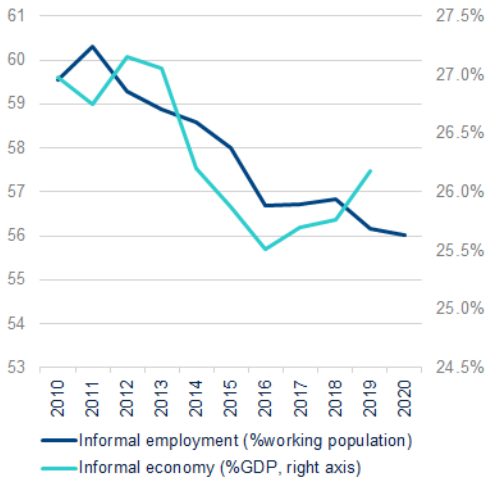
Programs like CoDi become relevant in an economy like Mexico with a widespread use of cash, and where 56 out of 100 workers belong to the informal labor market. This high figure places Mexico behind Chile, Argentina and Colombia, in terms of informality rate (Figure A1.). According to official data, in 2019 the informal economy in the country made up 26.2% of GDP (Figure A2.). Although the informality rate has decreased over the last decade (from 59.5% in 2010 to 56.0% in 2020 as a share of the working population), it has shown mixed performance throughout those 10 years. It is worth paying attention to the 2016–2018 period, which shows stagnation in the downward trend of informality, matching the growing trend in the size of the informal economy during those same years.

Figure A1. **INFORMALITY** (% WORKING POPULATION)



Source: INEGI

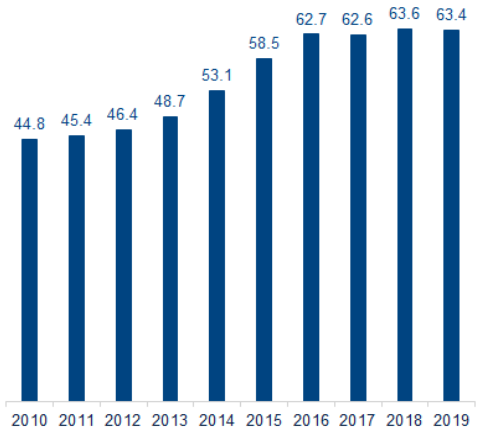
Figure A2. **INFORMAL EMPLOYMENT AND ECONOMY** (% WORKING POPULATION AND % GDP RESPECTIVELY)



Source: INEGI

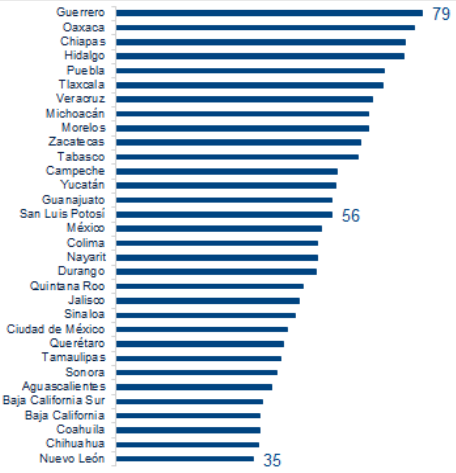
Together with the deepening of the informal economy during 2016–2018, the amount of banknotes and coins held by the public reached 62.7% of GDP (from 58.5% in 2015), signaling increased use of cash. According to the National Survey of Financial Inclusion 2018 (NSFI 2018), more than 90% of commercial transactions (retail, rent, public transportation, utilities and private services) were paid for in cash during that year. The same data shows that nearly 95% of purchases of MXN 500 or less (USD 25 or less) were paid for in cash, while 87% of purchases of MXN 501 or more use cash as the main payment method.

Figure A3. **BANKNOTES AND COINS HELD BY THE PUBLIC** (% GDP)



Source: Banxico

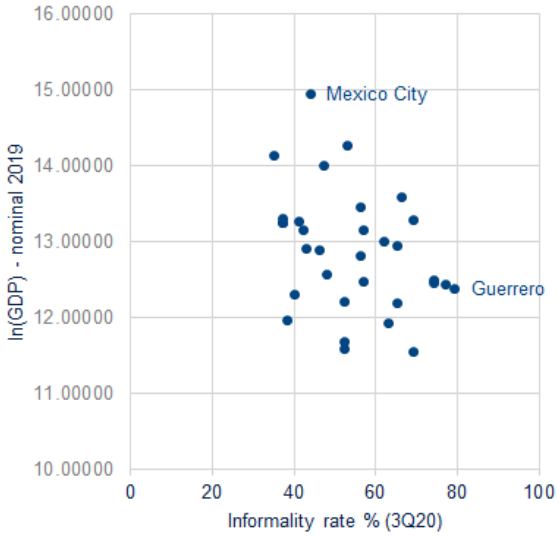
Figure A4. **INFORMALITY BY STATE** (% WORKING POPULATION)



Source: INEGI

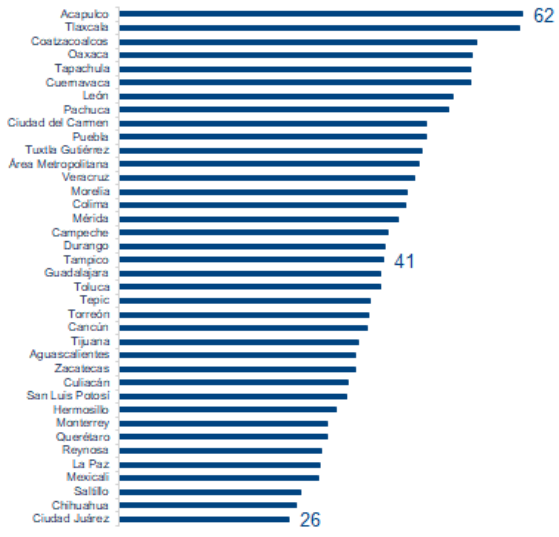
The states with the lowest incomes have the highest rates of informality in the country. The state of Guerrero reports that 79% of its working population work in informal conditions, followed by Oaxaca (77%) and Chiapas (74%). At the other end of the scale are Nuevo León y Chihuahua, with informality rates of 35% and 37%, respectively. The same geographic pattern is observed among the main cities of the country, with the urban centers with the lowest incomes reporting the highest levels of informality (Acapulco 62%, Tlaxcala 61%). Overall, the average rate of informality for the main cities in the country is 41%, 13 pp below the national figure (54%), which points to informality being concentrated in less urbanized areas.

Figure A5. **INFORMALITY VS. GDP, BY STATE** (% AND NATURAL LOGARITHM, RESPECTIVELY)



Source: INEGI

Figure A6. **INFORMALITY BY MAIN CITIES** (% WORKING POPULATION)



Source: INEGI

Annex 2 - Shap share coefficients

Table A1. **SHAP SHARE COEFFICIENTS FOR CARD SPENDING MODELS**

	Random Forest	Gradient Boosting	Decision Tree	Logit	OLS
North area	0.0749	0.0142	0.0128	0.002	0.019
South area	0.082	0.0102	0.0158	0.0001	0.0074
Metropolitan area	0.265	0.341	0.3792	0.0238	0.0969
Urban area	0.0925	0.0296	0	0.006	0.0419
Income	0.1473	0.257	0.2469	0.8703	0.7858
Deposit contracts	0.1082	0.0705	0.0686	0.0009	0.0024
Credit contracts	0.2301	0.2776	0.2767	0.0969	0.0467

Source: own calculations. BBVA and Clarity

Table A2. **SHAP SHARE COEFFICIENTS FOR CASH SPENDING MODELS**

	Random Forest	Gradient Boosting	Decision Tree	Logit	OLS
North area	0.0704	0.0474	0.0149	0.0022	0.0008
South area	0.0448	0.0404	0.0394	0.0001	0.0023
Metropolitan area	0.0534	0.0031	0.0024	0.0065	0.003
Urban area	0.0316	0.0393	0.0022	0.001	0.0066
Income	0.5118	0.6685	0.6941	0.9529	0.9392
Deposit contracts	0.0886	0.0372	0.0305	0.0002	0.0026
Credit contracts	0.1994	0.1641	0.2165	0.037	0.0456

Source: own calculations. BBVA and Clarity

Working Papers

2021

21/05 **Jaime Oliver, Álvaro Ortiz, Tomasa Rodrigo Saidé Salazar and Ignacio Tamarit:**
ESP/ Patrones de Consumo de Efectivo vs Tarjeta en México: una aproximación Big Data.
ING / Cash Vs Card Consumption Patterns in Mexico: A Machine Learning Approach.

21/04 **Ángel de la Fuente:** La financiación autonómica en 2020: una primera aproximación y una propuesta de cara a 2021.

21/03 **Ángel de la Fuente:** Las finanzas autonómicas en 2020 y entre 2003 y 2020.

21/02 **Joxe Mari Barrutiabengoa, J. Julián Cubero and Rodolfo Méndez-Marcano:** Output-side GHG Emissions Intensity: A consistent international indicator.

21/01 **Ángel de la Fuente y Pep Ruiz:** Series largas de VAB y empleo regional por sectores, 1955-2019 Actualización de *RegData-Sect* hasta 2019.

2020

20/17 **Amparo Castelló-Climent and R. Doménech:** Human Capital and Income Inequality Revisited.

20/16 **J.E. Boscá, R. Doménech, J. Ferri, J.R. García and C. Ulloa:** The Stabilizing Effects of Economic Policies in Spain in Times of COVID-19.

20/15 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2018.

20/14 **Ángel de la Fuente:** El impacto de la crisis del Covid sobre el PIB de las CCAA en 2020: una primera aproximación.

20/13 **Ali B. Barlas, Seda Guler Mert, Álvaro Ortiz and Tomasa Rodrigo:** Investment in Real Time and High Definition: A Big Data Approach.

20/12 **Félix Lores, Pep Ruiz, Angie Suárez y Alfonso Ugarte:** Modelo de precios de la vivienda en España. Una perspectiva regional.

20/11 **Ángel de la Fuente:** Series largas de algunos agregados económicos y demográficos regionales: Actualización de RegData hasta 2019.

20/10 **Ángel de la Fuente:** La liquidación de 2018 del sistema de financiación de las comunidades autónomas de régimen común.

20/09 **Lucía Pacheco Rodríguez and Pablo Urbiola Ortún:** From FinTech to BigTech: an evolving regulatory response.

20/08 **Federico D. Forte:** Network Topology of the Argentine Interbank Money Market.

20/07 **Ángel de la Fuente:** Las finanzas autonómicas en 2019 y entre 2003 y 2019.

20/06 **Vasco M. Carvalho, Juan R. Garcia, Stephen Hansen, Álvaro Ortiz, Tomasa Rodrigo, José V. Rodríguez Mora and Pep Ruiz:** Tracking the COVID-19 Crisis with High-Resolution Transaction Data.

20/05 **Jinyue Dong and Le Xia:** Forecasting modeling for China's inflation.

20/04 **Le Xia:** Lessons from China's past banking bailouts.

20/03 **Ángel de la Fuente y Pep Ruiz:** Series largas de VAB y empleo regional por sectores, 1955-2018. RegData_Sect FEDEA-BBVA (v5.0_1955-2018).

20/02 **Luis Antonio Espinosa y Juan José Li Ng:**
ESP/ El riesgo del sargazo para la economía y turismo de Quintana Roo y México.
ING / The risk of sargassum to the economy and tourism of Quintana Roo and Mexico.

20/01 **Ángel de la Fuente:** La dinámica territorial de la renta en España, 1955-2018. Los determinantes directos de la renta relativa: productividad, ocupación y demografía.

CLICK HERE TO ACCESS THE WORKING DOCUMENTS PUBLISHED IN
Spanish and English

DISCLAIMER

The present document does not constitute an “Investment Recommendation”, as defined in Regulation (EU) No 596/2014 of the European Parliament and of the Council of 16 April 2014 on market abuse (“MAR”). In particular, this document does not constitute “Investment Research” nor “Marketing Material”, for the purposes of article 36 of the Regulation (EU) 2017/565 of 25 April 2016 supplementing Directive 2014/65/EU of the European Parliament and of the Council as regards organisational requirements and operating conditions for investment firms and defined terms for the purposes of that Directive (MIFID II).

Readers should be aware that under no circumstances should they base their investment decisions on the information contained in this document. Those persons or entities offering investment products to these potential investors are legally required to provide the information needed for them to take an appropriate investment decision.

This document has been prepared by BBVA Research Department. It is provided for information purposes only and expresses data or opinions regarding the date of issue of the report, prepared by BBVA or obtained from or based on sources we consider to be reliable, and have not been independently verified by BBVA. Therefore, BBVA offers no warranty, either express or implicit, regarding its accuracy, integrity or correctness.

This document and its contents are subject to changes without prior notice depending on variables such as the economic context or market fluctuations. BBVA is not responsible for updating these contents or for giving notice of such changes.

BBVA accepts no liability for any loss, direct or indirect, that may result from the use of this document or its contents.

This document and its contents do not constitute an offer, invitation or solicitation to purchase, divest or enter into any interest in financial assets or instruments. Neither shall this document nor its contents form the basis of any contract, commitment or decision of any kind.

The content of this document is protected by intellectual property laws. Reproduction, transformation, distribution, public communication, making available, extraction, reuse, forwarding or use of any nature by any means or process is prohibited, except in cases where it is legally permitted or expressly authorised by BBVA.

ENQUIRIES TO:
BBVA Research: Azul Street, 4. La Vela Building – 4th and 5th floor. 28050 Madrid (Spain).
Tel. +34 91 374 60 00 y +34 91 537 70 00 / Fax (+34) 91 374 25
bbvaresearch@bbva.com www.bbvaresearch.com