National Accounts in a World of Naturally Occurring Data: An Application to Consumption

Gergely Buda
Barcelona GSE

Vasco Carvalho
University of Cambridge

Stephen Hansen
Imperial College

Alvaro Ortiz
BBVA Research

Tomasa Rodrigo
BBVA Research

Sevi Rodriguez Mora
University of Edinburgh

June 15, 2021
** Work in Progress! **
Financial Transaction Data

Pre-COVID: growing interest, relatively small-scale datasets.

COVID: explosion of papers that use payments data for tracking activity.

Post-COVID: use of new data sources for national accounts construction?
Financial Transaction Data

Pre-COVID: growing interest, relatively small-scale datasets.

COVID: explosion of papers that use payments data for tracking activity.

Post-COVID: use of new data sources for national accounts construction?

Our project takes near-complete record of financial transactions from BBVA and asks whether they can be used to build national statistics.
Why Transaction Data?

Available in real time →
tracking impact of shocks on national accounts.

Much larger samples than in traditional surveys →
ability to obtain more granular measures (e.g. geography, income)

Collection of data is cheaper for public sector (although private sector bears significant costs) →
particularly relevant in lower-income countries.

Main challenge is that data collection is “accidental” and has many potential sources of bias and noise.
Banco Bilbao Vizcaya Argentaria

Banco Bilbao Vizcaya Argentaria is Spanish multinational bank with operations in 30 countries.

Largest markets include Spain, Mexico, and Turkey.

Top-50 global bank by total assets, second largest in Spain (727B EUR).

Partnership with Big Data division of BBVA Research.
Today we focus mainly on card data:

1. Relatively well organized

2. By far the most common type of large-scale payments data

Raw spending is biased and noisy; straightforward filters yield good measure of final household consumption.

Application to distributional national accounts via account-level link between spending and income.
Results on Consumption

GDP Household consumption vs. transaction data: Y-o-Y quarterly growth

- Naturally occurring transactions: baseline data
- Naturally occurring transactions: lowest RMSE
- INE Final Household Consumption

Card Data Description

BBVA card transactions from 2015 Q2-2021 Q1.

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transactions</td>
<td>4,739,263,647</td>
</tr>
<tr>
<td>Total volume of transactions</td>
<td>254,862,281,174</td>
</tr>
<tr>
<td>Number of distinct customer IDs</td>
<td>7,224,844$^1$</td>
</tr>
<tr>
<td>Number of ‘Active Customers’$^2$</td>
<td>1,921,652</td>
</tr>
</tbody>
</table>

$^1$1,717,514 self-employed

$^2$At least one transaction in each quarter
Customers with transactions and sampling frame of active customers
Gender distribution

- Male: INE Spain Demographics
- Female: Customer Demographics in Naturally Occurring Data
Classifying Good Purchases

Each card transaction has metadata on the type of purchase.

Breakdown by online/offline.

Each transaction has merchant client code (MCC, 838 in total).

MCCs are mix of generic international codes and Spanish-specific codes for large retailers (Mercadona, Zara, etc).

There is also an MCC for cash withdrawals from ATMs.
Mapping MCC to COICOP

We attempt to assign each MCC to one of the twelve COICOP categories.

In case of multiproduct retailers (e.g. large department stores), we distribute purchases across COICOP by using auxiliary data sources.

33 MCCs are unclassified consumption (‘wholesale purchase’, ‘direct sales’).

38 do not refer to consumption (‘pawnshop’, ‘cryptocurrencies’).
Proportions of spending per COICOP, 2019

% of naturally occurring data volume in COICOPs % of Household spending in COICOPs (INE)

01 Food and non-alcoholic beverages
07 Transport
11 Restaurants and hotels
12 Miscellaneous goods and services
09 Recreation and culture
03 Clothing and footwear
05 Furnishings, household equipment and repair
06 Health
08 Communication
02 Alcoholic beverages and tobacco
10 Education
Aggregate Consumption Based on Raw Spending

Form raw consumption indicator by computing total spending per quarter:

1. Spending on all cards by all customers
2. Spending on all MCC except for cash
3. Online and offline spending

The resulting series mimics the spending measures that would arise without account information.
Aggregate Consumption Based on Filtered Spending

We have highlighted aspects of the raw data that potentially lead to bias and/or noise.

We propose various “switches” that act to transform or filter the raw data, ideally bringing aggregate spending in line with national accounts.

We optimize over these switches by minimizing MSE of growth in derived spending series and in official consumption series.

Goal is not curve fitting, but to document the importance of different factors in building national accounts measure.
Switches

1. Removal of non-consumption MCCs.
2. Include cash withdrawals.
3. Remove online transactions.
4. Active clients filter for alternative thresholds for minimum number of transactions per quarter: 1, 5, 10.
5. Exclude self-employed clients.
6. Weight by INE demographic characteristics.
Switches

1. Removal of non-consumption MCCs.
2. Include cash withdrawals.
3. Remove online transactions.
4. Active clients filter for alternative thresholds for minimum number of transactions per quarter: 1, 5, 10.
5. Exclude self-employed clients.
6. Weight by INE demographic characteristics.

Result: all switches turned on to minimize MSE, with active client threshold of 10.

Same is true in pre-COVID sample.
Transformed Spending vs Consumption: Entire Sample

GDP Household consumption vs. transaction data: Y-o-Y quarterly growth

- Naturally occurring transactions: baseline data
- Naturally occurring transactions: lowest RMSE
- INE Final Household Consumption
## Effect on MSE of Changing Switch from Optimum

<table>
<thead>
<tr>
<th></th>
<th>Whole Series</th>
<th>Pre-COVID Series</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>Δ RMSE</td>
</tr>
<tr>
<td>Optimal</td>
<td>0.0250</td>
<td>0</td>
</tr>
<tr>
<td>No Active Customer</td>
<td>0.1408</td>
<td>0.1158</td>
</tr>
<tr>
<td>Exclude Cash</td>
<td>0.0685</td>
<td>0.0435</td>
</tr>
<tr>
<td>Include Online</td>
<td>0.0272</td>
<td>0.0022</td>
</tr>
<tr>
<td>No Demographic Weighting</td>
<td>0.0264</td>
<td>0.0014</td>
</tr>
<tr>
<td>Non-consumption MCC</td>
<td>0.0254</td>
<td>0.0004</td>
</tr>
<tr>
<td>Raw Data</td>
<td>0.0835</td>
<td>0.0585</td>
</tr>
</tbody>
</table>

Non-Card Transaction Data
Spending and Income

We next construct distributional national accounts by pairing account-level spending and income.

Spending measure includes cash; excludes online; excludes non-consumption MCCs.

Three sources of income (available since 2017): salary, government benefits, pensions.

Sample restricted to active clients who receive at least one source of income in each month of 2017-2019.

Approximately 260,000 individuals.
Comparison of Income Distributions

Income distribution, naturally occurring data vs. INE statistics (pooled over 2017-2019)
Spending by Income Level

Card spending by income groups

Average yearly card spending (2017-2019)

Average yearly net income (2017-2019)

<10k
10-15K
15-20K
20-25K
25-30K
30-35K
35-40K
40-45K
45-50K
50K+
Spending by Income Percentile

Nominal Card Spending per Income Percentile

Average yearly spending per person (2017-2019)

Income percentiles (2017-2019)
Income and Spending Growth

Spending and income growth by percentile

Average yearly growth (2017-2019)

Income percentiles (2017-2019)
Income Growth over Time

Growth of income by percentile and year

Yearly growth of income

0.05
0.04
0.03
0.02
0.01
0.00
−0.01

Income percentiles (2017-2019)

2017 to 2018
2018 to 2019
2019 to 2020
Spending Growth over Time

Growth of spending by percentile and year
Conclusion

Our goal is to take near-complete record of financial transactions from one of world’s largest banks and construct national accounts.

Promising proof-of-concept exercise for consumption in Spain.

Exercise can be replicated with other banks in other countries.

Ongoing work for other components of national accounts.
Demographic Weighting

Let $x_{a,r,s}^{\text{INE}}$ be the number of Spanish adults:

- In age band $a$
- Living in region $r$
- With sex $s$

Let $x_{a,r,s}^{\text{BBVA}}$ be the number of BBVA customers.

Let $y_{a,r,s}$ be total spending of BBVA customers.

Both BBVA-derived variables are with respect to given switch.

Weighted spending is $y_{a,r,s}^{w} = y_{a,r,s} \times (x_{a,r,s}^{\text{INE}}/x_{a,r,s}^{\text{BBVA}})$. 

Return
Non-Card Data

We also have access to full set of non-card transactions which require additional filtering.

We have currently tabulated account-level spending on utility categories: water, electricity, phone, gas.

Spending allocated to date on which payment is taken from bank.

(Housing remains largest unaccounted-for component of consumption).
Naturally Occurring Data Total Utilities vs. INE 2019 Consumption Proportions

- **Non-Card Transactions: Electricity**
- **Non-Card Transactions: Gas**
- **Non-Card Transactions: Phone & Internet & TV**
- **Non-Card Transactions: Water**
- **INE Water 2019 annual %**
- **INE Electricity 2019 annual %**
- **INE Gas 2019 annual %**
- **INE Phone & Internet & TV 2019 annual %**
Y-o-Y quarterly growth, Including non-card utilities

Quarter-on-quarter % change

Quarters


Naturally occurring data with optimal switches - no utilities (RMSE:0.0234)
Naturally occurring data with optimal switches - including utilities (RMSE: 0.020)
INE Final Household Consumption

RMSE without utilities: 0.0235; RMSE with utilities: 0.0203.
Spending by Income Level (Log Scale)