




















National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption

Gergely Buda (BSE)
Vasco M. Carvalho (Cambridge)
Stephen Hansen (UCL)
Álvaro Ortiz (BBVA Research)
Tomaso Rodrigo (BBVA Research)
José V. Rodríguez Mora (Edinburgh)

July 19, 2023

- Introduction 
- Contributions 
- Building a Consumption Survey 
- Aggregate National Accounts 
- Distributional Accounts 
- Income and GRID 
- Individual Consumption Dynamics 
- Conclusion 

- Extra 
 - Short and Variable Lags 
 - Households and distribution 
 - Alternative Assumptions 
 - Growth divergence during COVID 
 - Good coverage across all COICOPS. 
 - Distribution across COICOPS and people 
 - Distributional by time and COICOP 
 - Distribution of Growth by COICOP 
 - Two perspectives of Growth and Inequality 
 - Dynamics Frequencies Coicops 

- Modern payment systems generate an enormous amount of real-time data on activity that is physically recoverable and has the potential to inform on the whereabouts of economic activity with unprecedented accuracy.
- Advantages of such data include timeliness, granularity, and cost to statistical agencies (albeit not to private sector).
- Growing interest in non-traditional data for tracking the economy, especially in the wake of COVID-19.
- Already long literature. Probably the closest in spirit is Anderson et al, albeit with different focus (cell disaggregation of NA, not panel construction)
- But few, if any, attempts in the academic literature to build extensive and encompassing substitutes of surveys and national accounting objects from first principles using large-scale payment data.

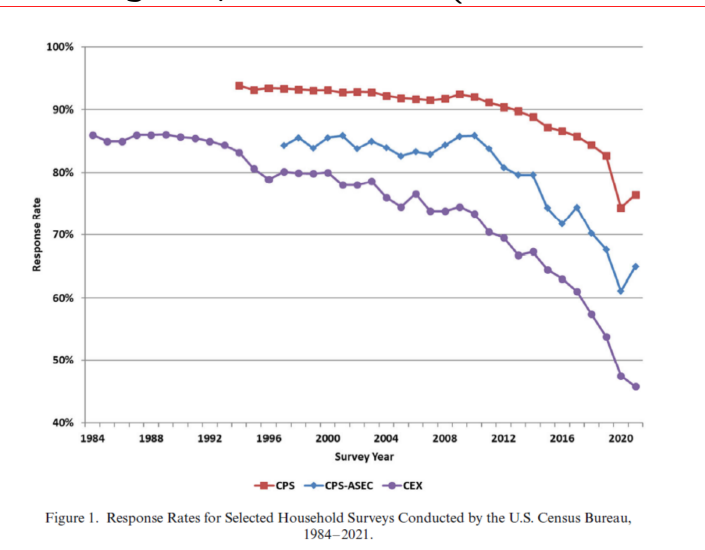
First proof of concept that **naturally occurring transaction data**, arising through the decentralized activity of millions of economic agents, can be **organized via national accounting rules** and then harnessed to produce a large-scale, high-quality and highly-detailed **consumption survey** that by simple aggregation reproduces National Accounting objects

- Universe of BBVA retail accounts in Spain by BBVA
- Allowing us to track expenditure as it flows out of these accounts, transaction by transaction
- 3 billion individual transactions by 1.8 million BBVA customers, from 2016 to 2021

Consumption Surveys

- Don't aggregate to national accounts consumption
- Under-reporting is not constant across income profiles
- Limited panel coverage
- Low frequency
- Declining response rates
- Difficult substitution/validation with administrative data.

- Decreasing response rate. (Abraham 2022)



National Accounts

Transaction Data can be incorporated into national accounting measures (Bean 2016).

- In many countries NA are sparse, or non-existent (Silungwe et al 22):
 - 33% of countries do not publish quarterly NA (50% in Africa)
 - Only 4 European and 5 Asian countries produce quarterly NA within 30 days of the reference period.
 - 25% of countries have no Household Budget Survey.
- In lower-income countries, transaction data may be the *only* reliable source of information for building national accounts.
- Measurement may be biased by political reasons (Martinez JPE 2022).
- Democratization of National Accounts looks like a good idea.

1. Show how to construct representative panel of household expenditure. Massive survey.

- Including all forms of expenditures: cards, direct debits, transfers, cash...
- Categorize transactions across harmonized consumption spending categories
- Filter out non-consumption expenditures (transfers to saving accounts, household-to-household transfers or tax payments)
- Impute consumption of housing services for all households
- Construct large sampling frame of households that is representative along demographic observables (gender, age and spatial cells) so as to mimic the characteristics of the Spanish adult population.

1. Show how to construct representative panel of household expenditure. Massive survey.
2. Show that it aggregates to Quaterly National Accounts

- "Gasto en Consumo Final de los Hogares", produced quarterly by INE.
- Very good match in spite of vastly different methodology.
- Levels and Growth!

1. Show how to construct representative panel of household expenditure. Massive survey.
2. Show that it aggregates to Quaterly National Accounts
3. Create Distributional National Accounts for Consumption

- Macro-consistent, micro distribution of consumption.
- Description of Inequality in Consumption
 - Different Demographics
- Comparison to Encuesta de Presupuestos Familiares.
 - Right tail differences.
 - Advantage of Macro Aggregation.
- Distribution of Growth of Consumption and Inequality
 - COVID

1. Show how to construct representative panel of household expenditure. Massive survey.
2. Show that it aggregates to Quaterly National Accounts
3. Create Distributional National Accounts for Consumption
4. Study micro-structure of Consumption Dynamics.

- Lumpy Structure of Consumption Growth at individual level.
- Consumption growth difficult to approximate with Gaussian Distribution
- High skewness
- Excess Kurtosis (Thick Tails)

- Two problems for translating transaction data into a representative sample of the consumption of the population:

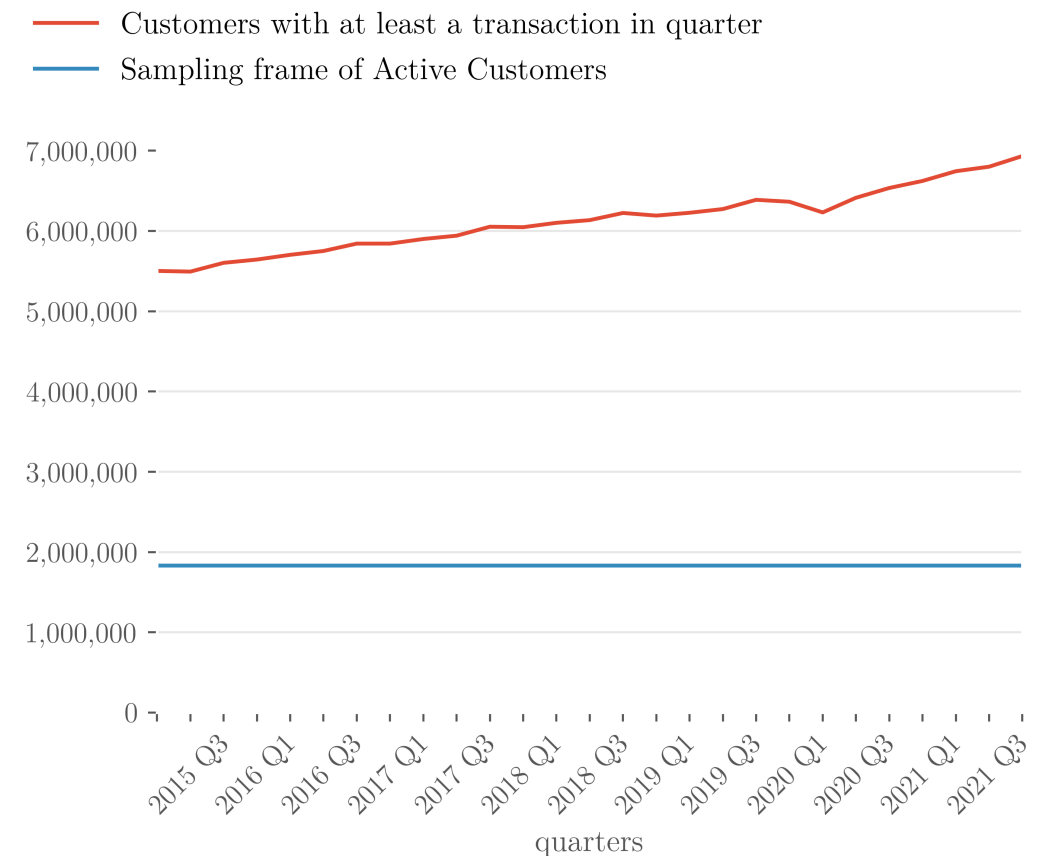
The client pool of a bank is not a representation of the population: biases

Spending is not the same than consumption.

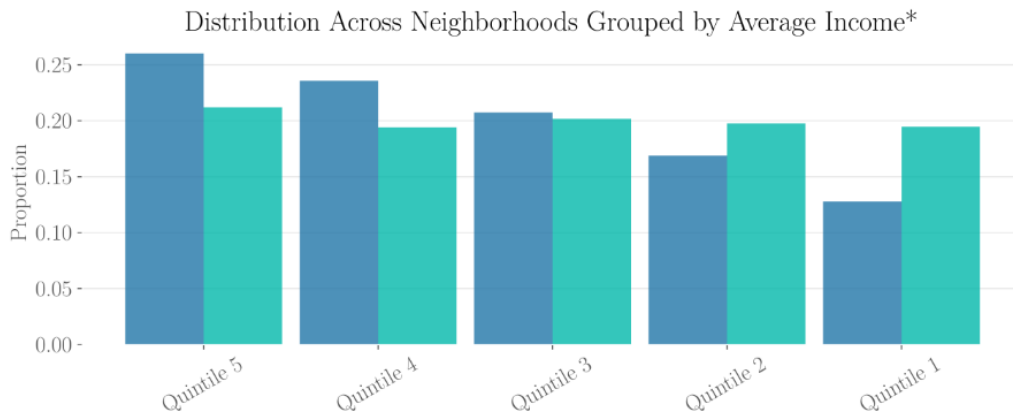
- Many movements out of a private account are not consumption: **financial movements, taxes...**
- Some consumption does not appear in any obvious manner as spending from an account: **housing**

Sample Frame

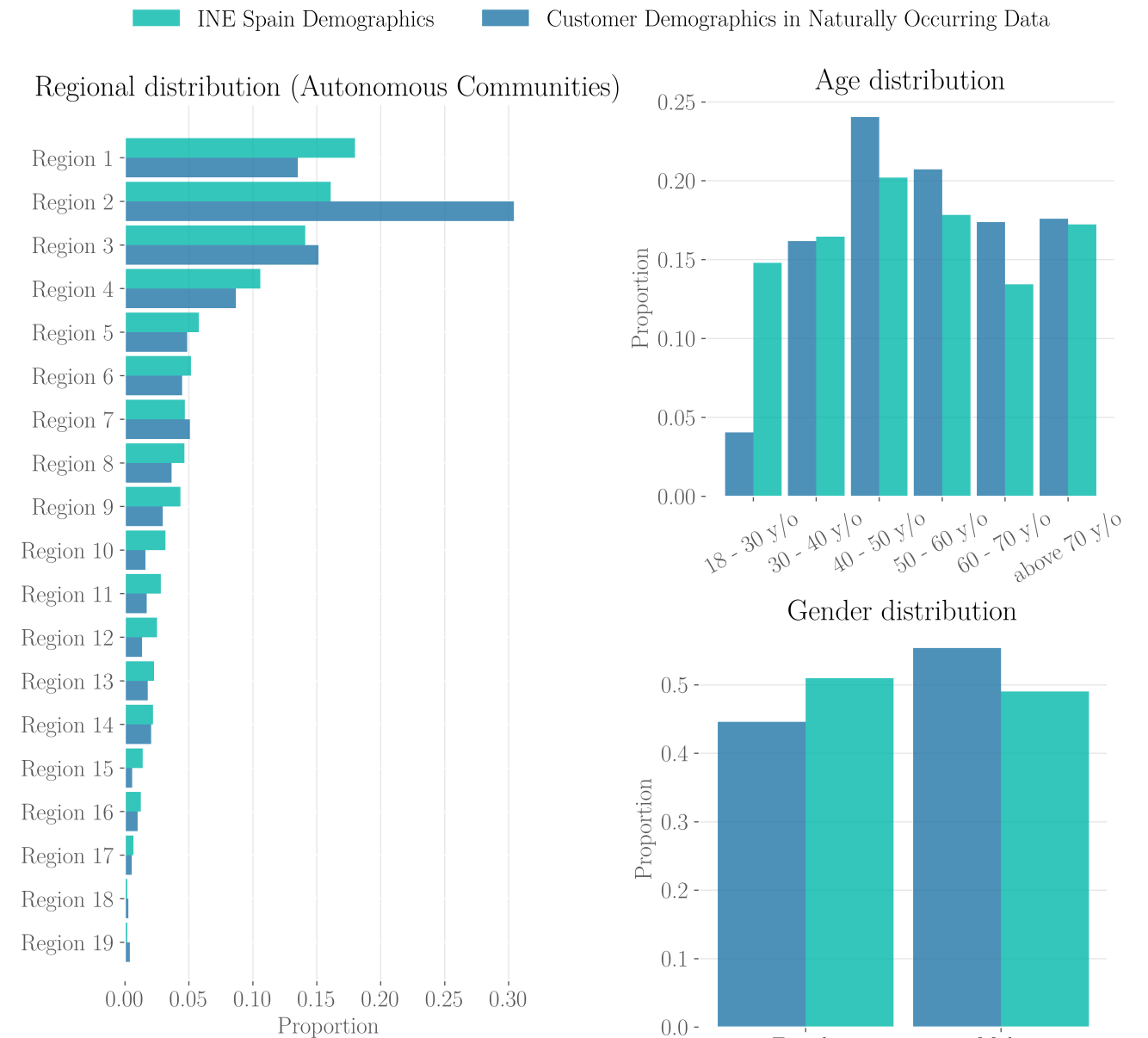
- 10,270,041 unique customers (2015-2021)
- Most spend infrequently or for short periods only.
- Define **“Active Customers”** as making at least 10 consumption related transactions in each quarter.
- 1,827,866



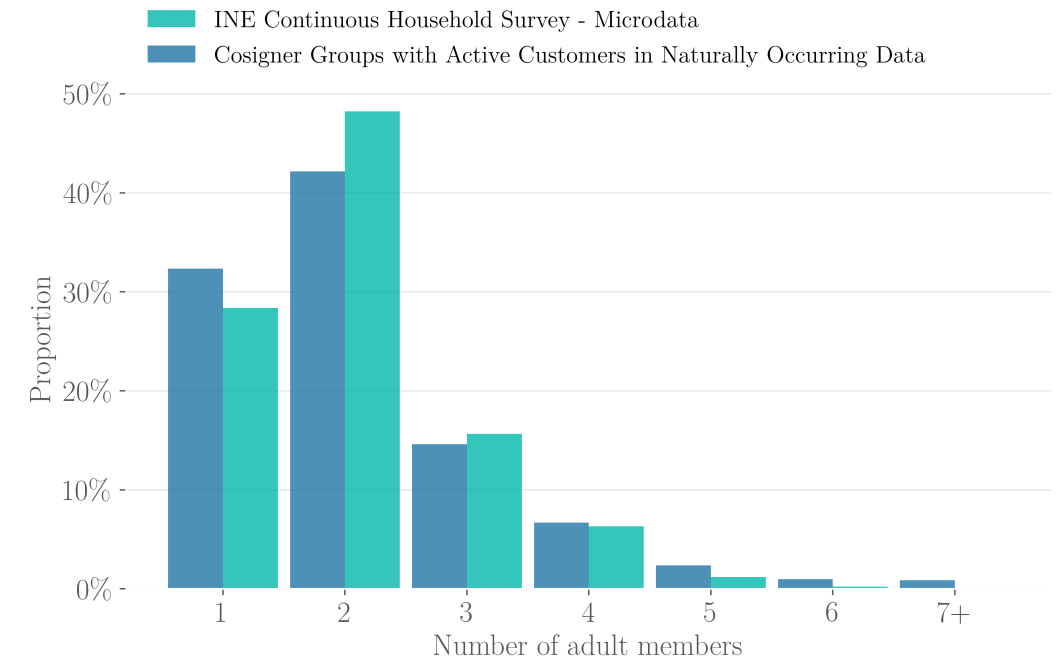
Demographics of Active Customers



*Excluding Basque Country, Navarra, Ceuta and Melilla



- Link clients into perceived household groups.
- Individuals with whom they share a contract and live in same postal code.
- We add married couple if not in sample.
- 1,589,280 household groups



	HBS 2016	HBS 2017	HBS 2018	HBS 2019	HBS 2020	BBVA Sample
Households	22,011	22,043	21,395	20,817	19,170	1,589,280
Adults	47,420	47,055	45,328	43,988	40,285	1,827,866

Clasification of Non-Housing Consumption Spending

- If a transaction is explicitly categorized in one of the 12 COICOPS.
- Follow national accounting principles wherever possible

Card data:

- Merchant Client Code (MCC) of the counterparty firm.
- Manual Mapping to COICOPS
- Multi-product retailers. Assigned by external data on distributions.

Direct Debit.

- ~ 100 internal labels.
- Manual Mapping
- When this is unclear, we read field, determine firm and use either MCC (if possible) or NACE code of firm to assign COICOP.

Transfers:

- String match counter-party name to commercial registry.
- If counter-party is located as a firm, we assign as above.

Cash Withdrawals.

- Both cash and over the counter.
- Assume is consumption.
- Assumptions on distribution.

Spending Category	Volume of Transactions	Number of Transactions
Offline Card Transactions	60,319 million	1,772 million
Online Card Transactions	11,858 million	313 million
Direct Debits	66,036 million	752 million
Cash Withdrawal	64,592 million	359 million
Transfers excl. rent	11,148 million	15 million

Determination Housing Consumption Spending

- We locate payment of rental for housing services.
 - Reading of free-text field in direct debits and transfers.
 - Minimum 100 EUR
 - Exclude parking, etc.
 - Payments made in 70 months.
 - 32,127 households.
- Use household covariates to predict monthly rent
 - Income (from BBVA table, six month average)
 - Utility Payments (direct debits)
 - Geography: 327 regions (consolidating postal codes)

Variable	Model	Test set
Spending on House Utilities	0.0884 (0.0008)	
Income	0.0362 (0.0011)	
N of Contract Groups	16,977	15,512
N of Observations	1,134,735	15,512
R^2	0.3911	
Adjusted R^2	0.3765	
Within R^2	0.1200	
Root MSE	204.6144	221.64

Out-of-sample behavior is reasonable with households that are 50-70 months in data.

Use covariates to IMPUTE housing consumption for the rest of the households (the vast majority)

Weighting and Sampling

- We observe the spending not consumption within household
- ASSUME equal spending among active clients within households, and half the weight of non active clients.

- Define cells of gender, age and region.
- Adjust demographic weight of cells to make them representative:

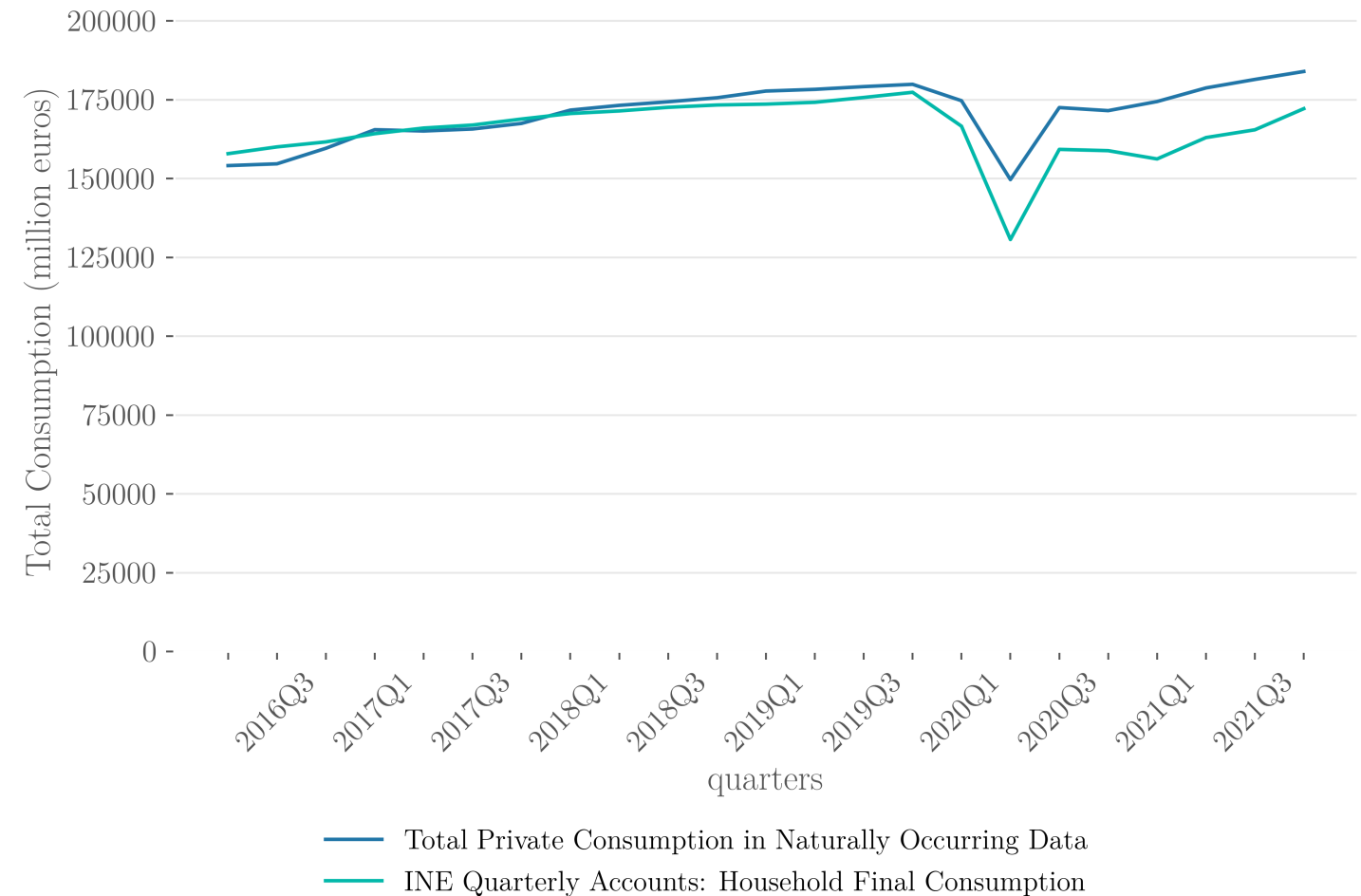
$$c_{g,a,r} = c_{g,a,r}^{\text{hh}} \left(\frac{x_{g,a,r}^{\text{INE}}}{x_{g,a,r}^{\text{BBVA}}} \right)$$

- In occasion we need to create a complete national sample.
- We draw $x_{g,a,r}^{\text{INE}}$ times from the pool of active client IDs within cell
- Sampling with replacement.

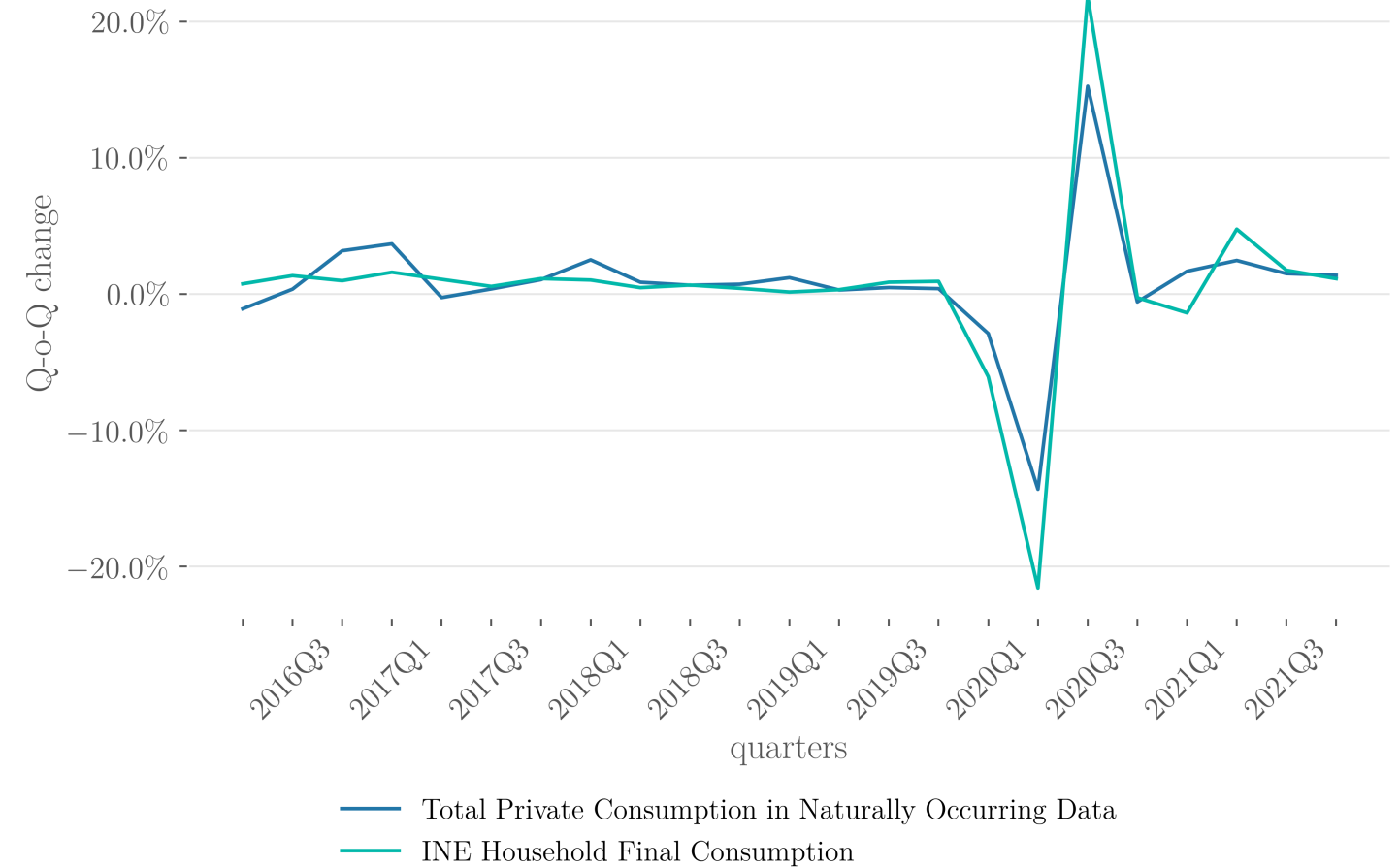
Aggregation into National Accounts

- Simple aggregation of data at quarterly frequency reproduces National Accounts
- Overall distribution across categories matches surveys.
- Arbitrary Frequency

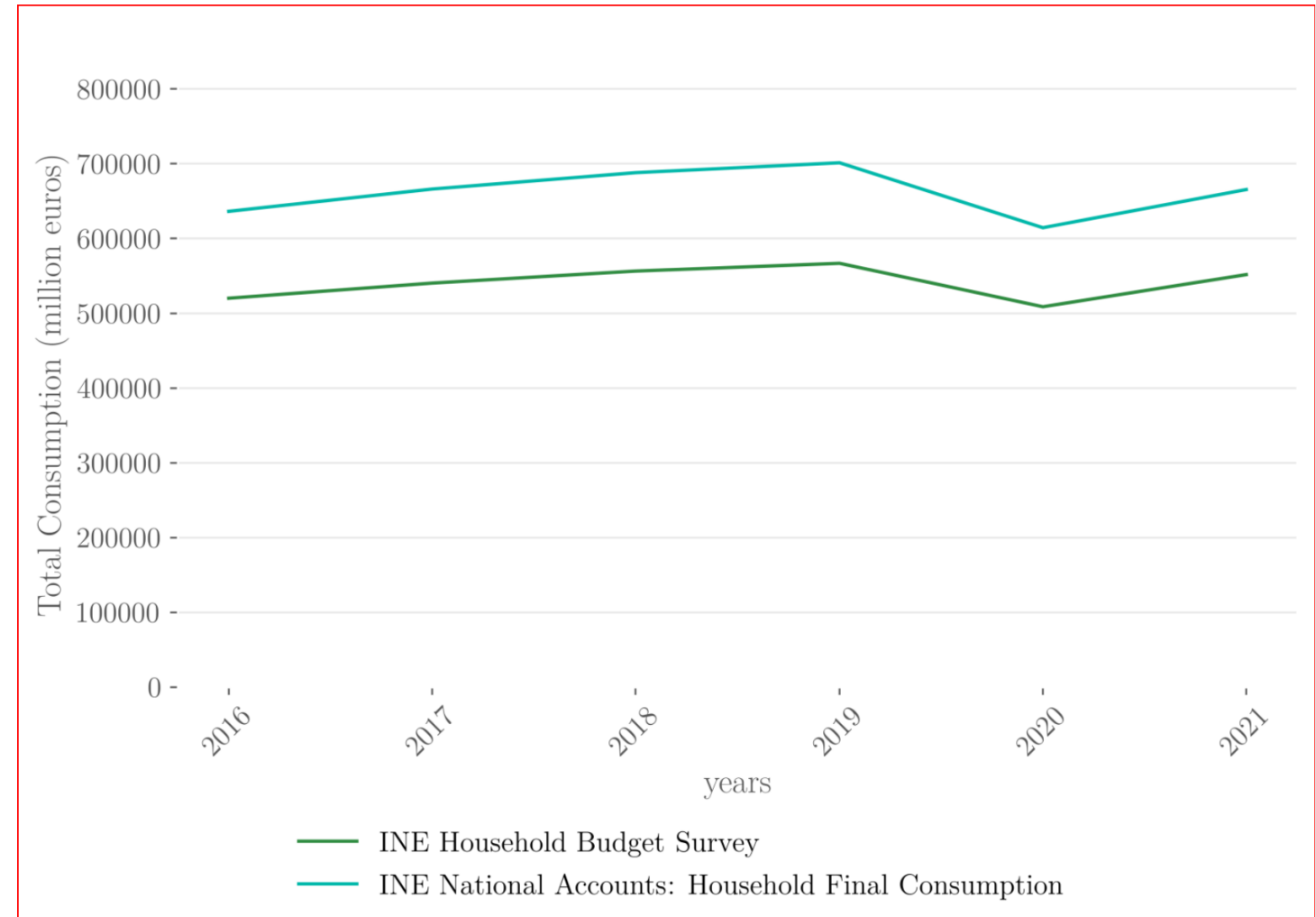
- Remarkable Similarity of levels.
- Even if accounted from a vastly different methodology
- Quarterly for equal comparison, but frequency could be even daily.



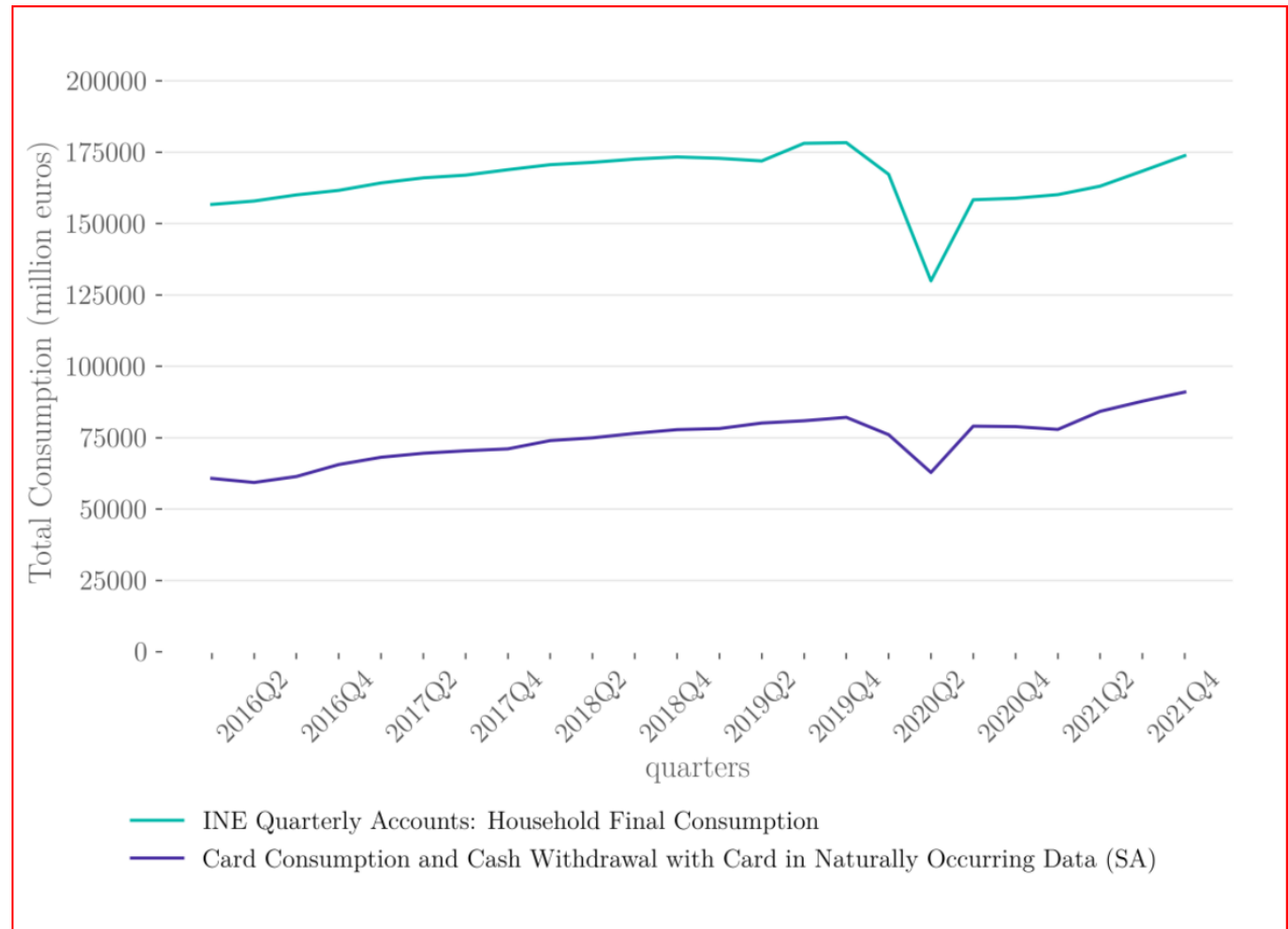
- Growth rates also track quite closely
- Just difference in volatility during COVID.



- This stands in contrast with the fact that the best available survey (HBS) undercounts consumption.
- General problem of surveys
- Our coverage is substantially better than surveys COICOP to COICOP.

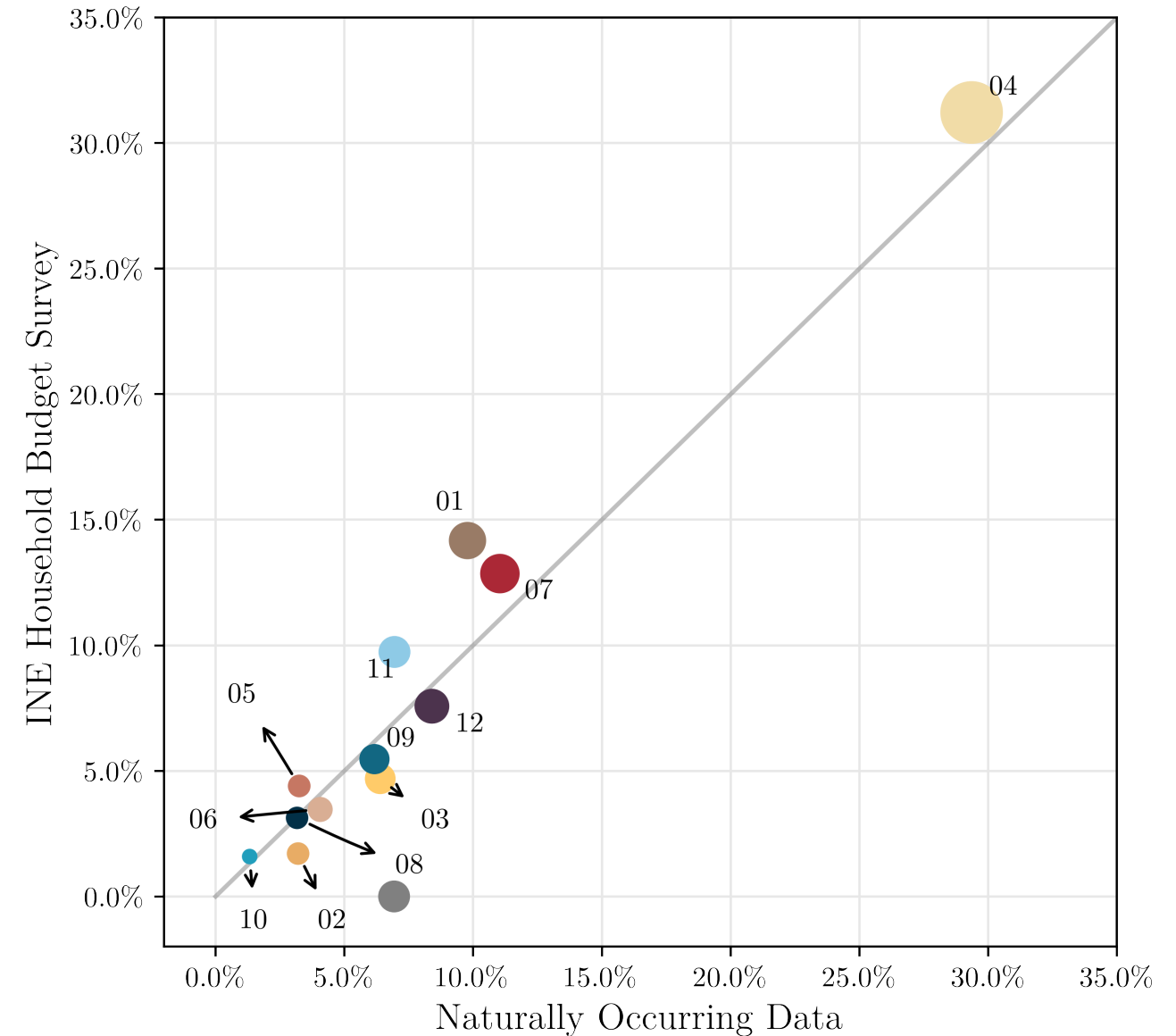


- Including only cards and cash account to about half spending.

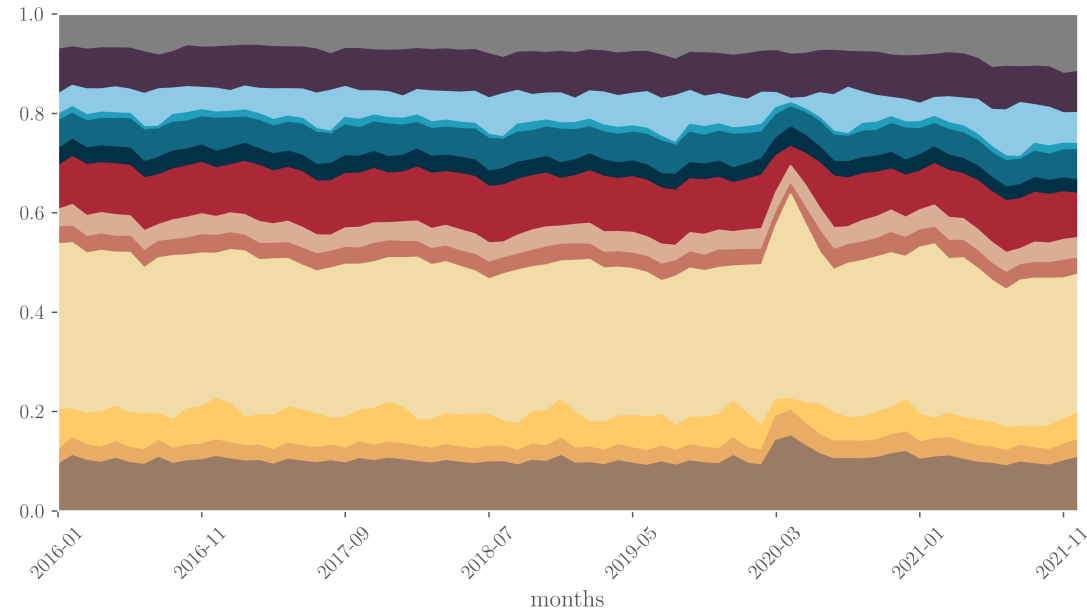


Good matching of distribution across COICOPS from HBS and national accounts.

- Cash is assumed to be consumed like offline cards.
- Adjusting per percentile of consumption

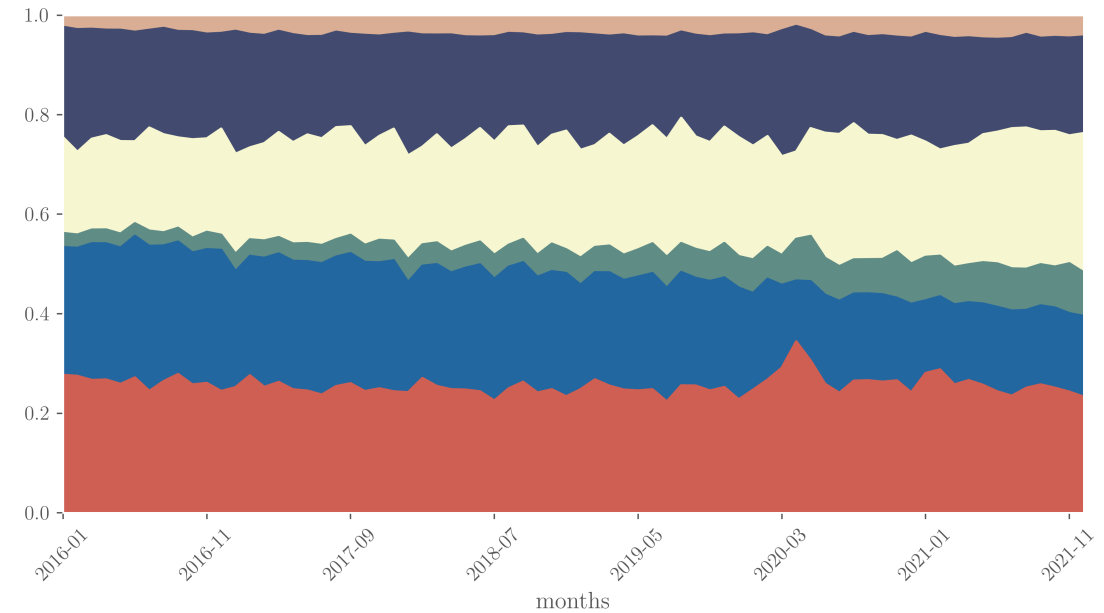


Very dense and rich time series



- | | |
|--|-------------------------------------|
| 01 Food and Non-Alcoholic Beverages | 08 Communication |
| 02 Alcoholic Beverages, Tobacco, and Narcotics | 09 Recreation and Culture |
| 03 Clothing and Footwear | 10 Education |
| 04 Housing, Water, Electricity, Gas, and Other Fuels | 11 Restaurants and Hotels |
| 05 Furnishings, Household Equipment and Maintenance | 12 Miscellaneous Goods and Services |
| 06 Health | Uncategorized |
| 07 Transport | |

COICOPS



- | | | |
|--------------|--------------|---------------|
| Imputed Rent | Online Card | Direct Debit |
| Cash | Offline Card | Wire Transfer |

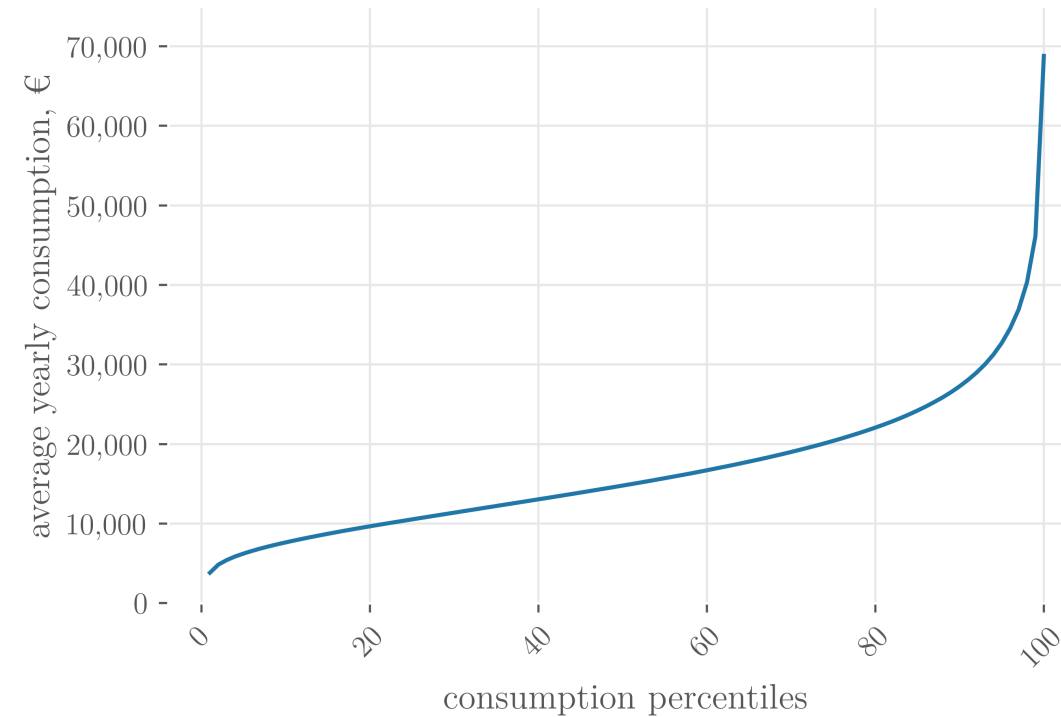
Means of Payment

- Real series can be calculated: Price data available at *Month* \times *Province* \times *Category*
- Nominal data frequency can go up to daily.
- Cuts can be made arbitrarily:
 - Region, town, neighborhood
- It allows to look at distributional issues with detail.

Cross-sectional Inequality of Consumption

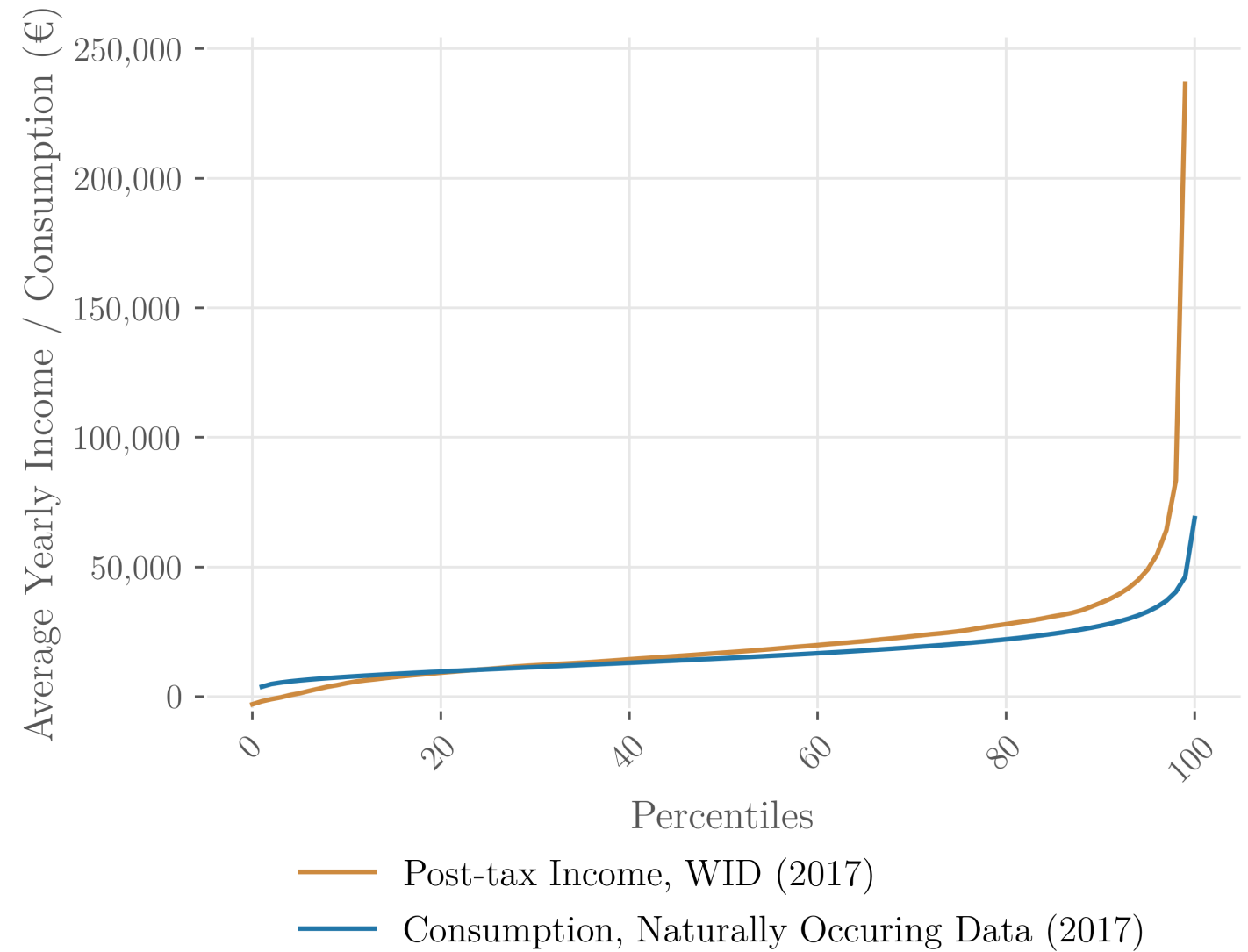
- The data allows to create Distributional national accounts
 - Their aggregation reproduces National Accounts
- ... while one can study distributional aspects.

- Macro-consistent, distribution of Consumption.
- It aggregates into NA
- Distributional Accounts directly from data.
- **No imputation.**



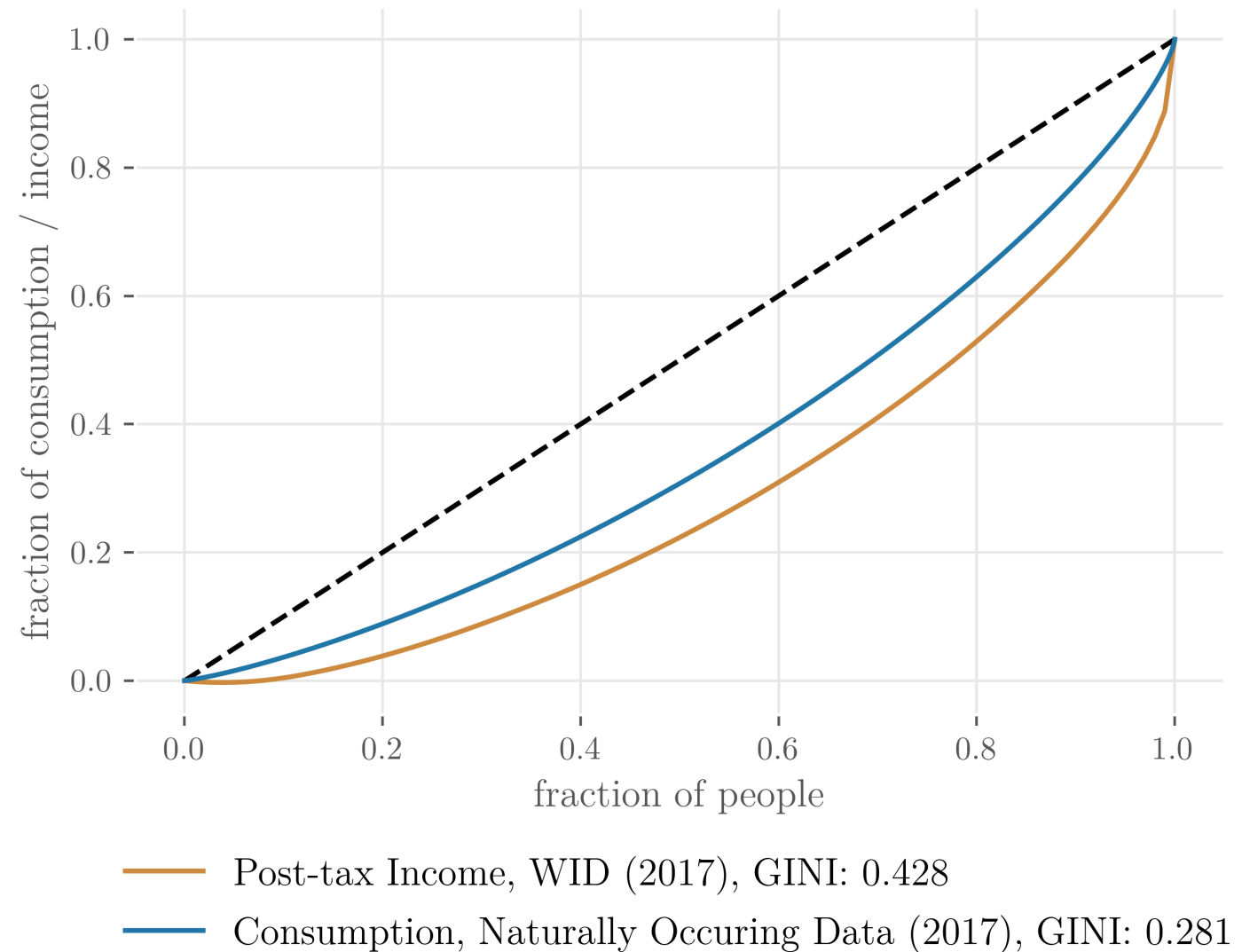
Distribution of Consumption, 2019

Comparison Distribution of Income (WID) and Consumption.



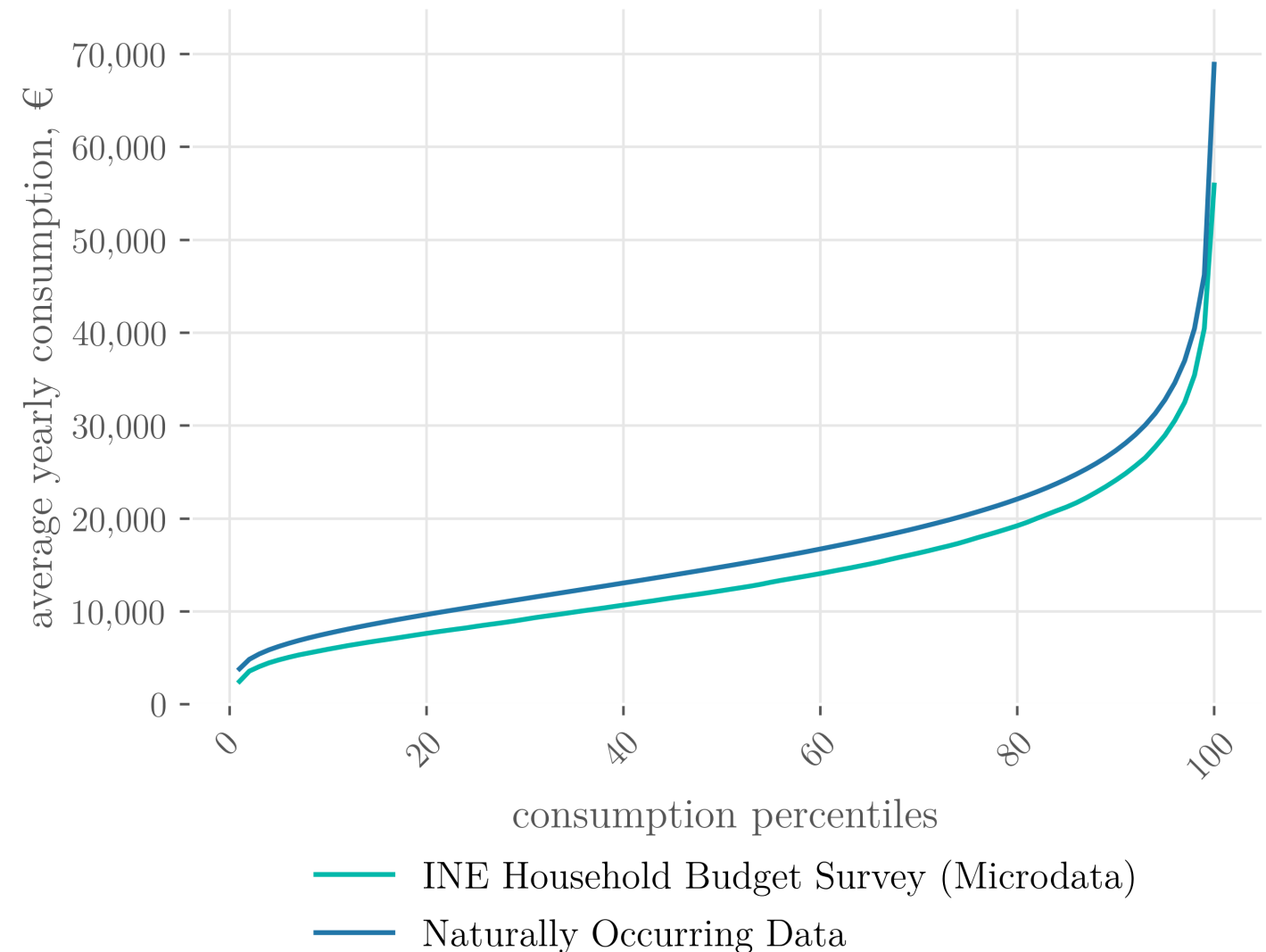
Comparison Distribution of Income (WID) and Consumption.

- As it should be expected, less inequality of consumption.



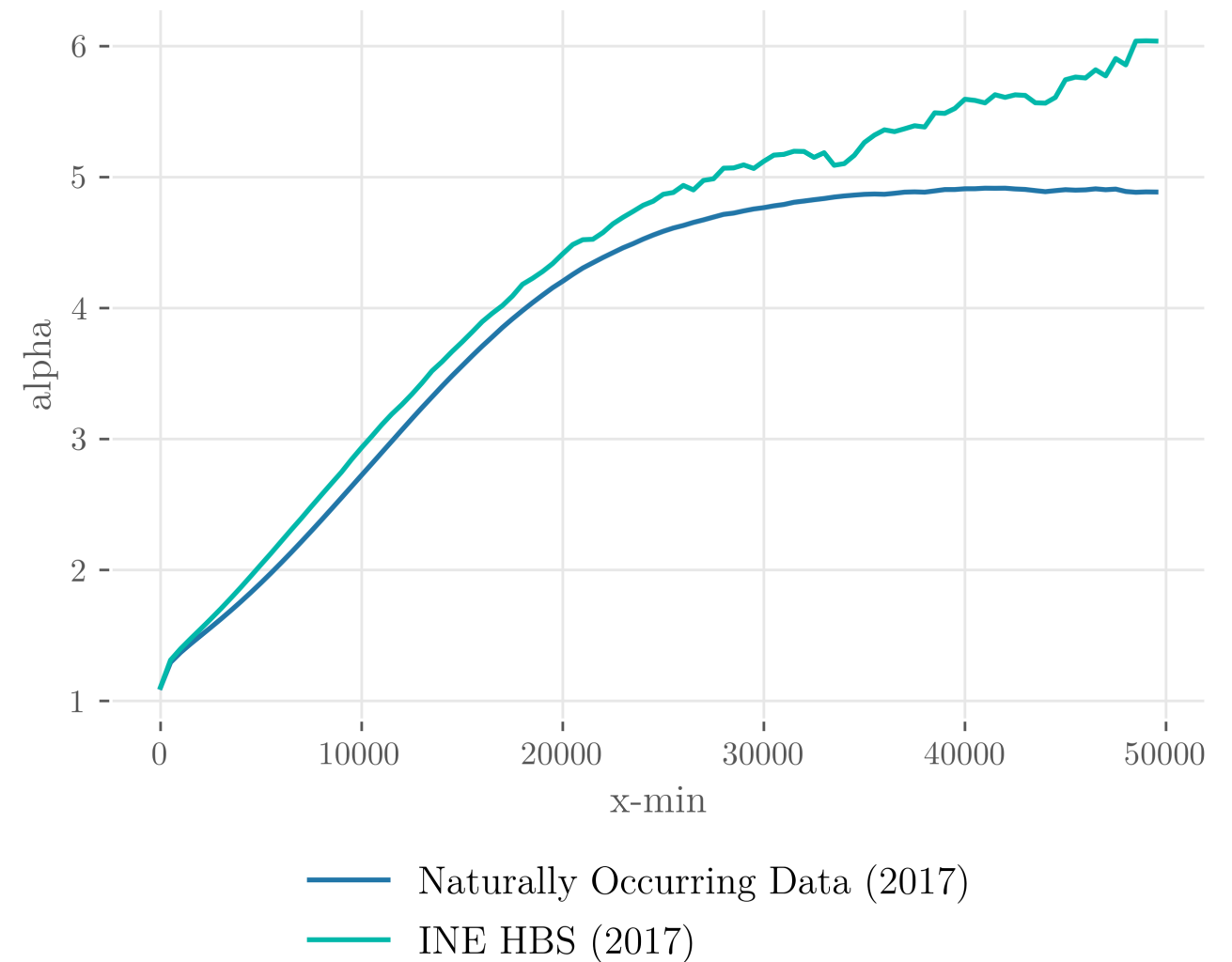
Comparison of Distribution of consumption with HBS.

- Remember: difference in level.
- Right tail difference: Bigger share of consumption among the people who consumes most.



Comparison of Distribution of consumption with HBS.

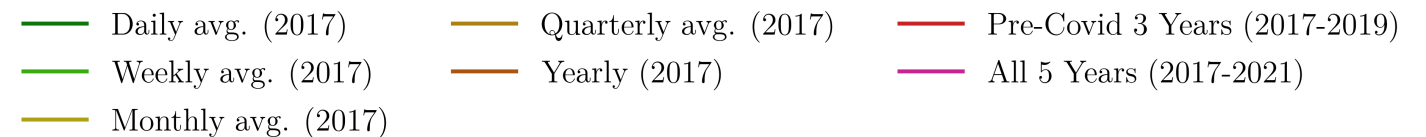
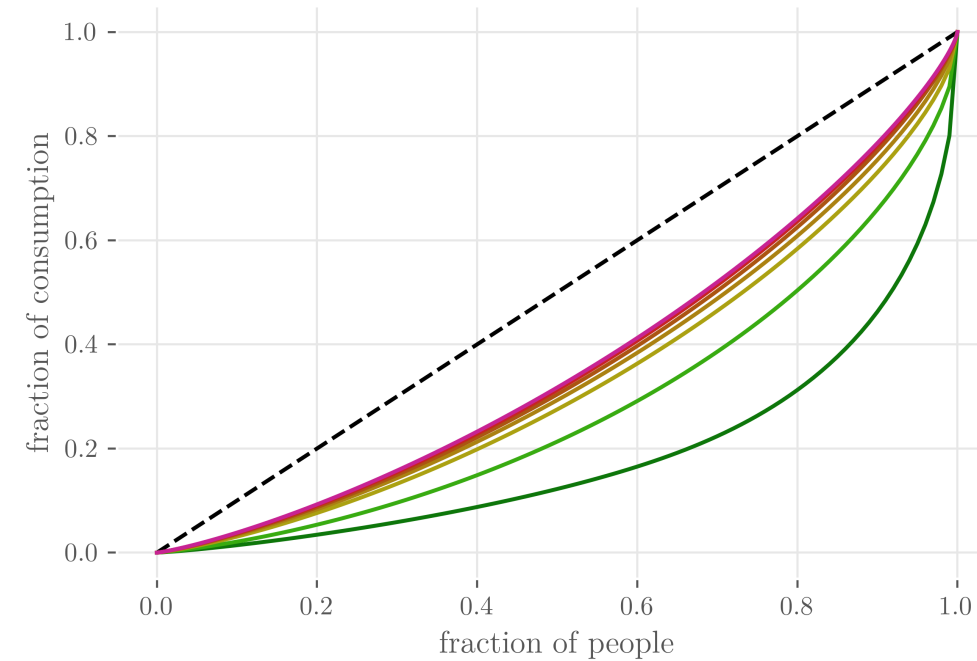
- Naturally occurring data: Thick tail. Power Law. (like in Income distribution)



- Across Time frequencies
- Consumption is a flow.
- Unite of Time aggregation matters.
- Lumpy.
- Critical in survey design.

Frequency	Gini index
Daily avg. (2017)*	0.629
Weekly avg. (2017)	0.439
Monthly avg. (2017)	0.338
Quarterly avg. (2017)	0.307
Yearly (2017)	0.281
Pre-Covid 3 Years (2017-2019)	0.273
All 5 Years (2017-2021)	0.265

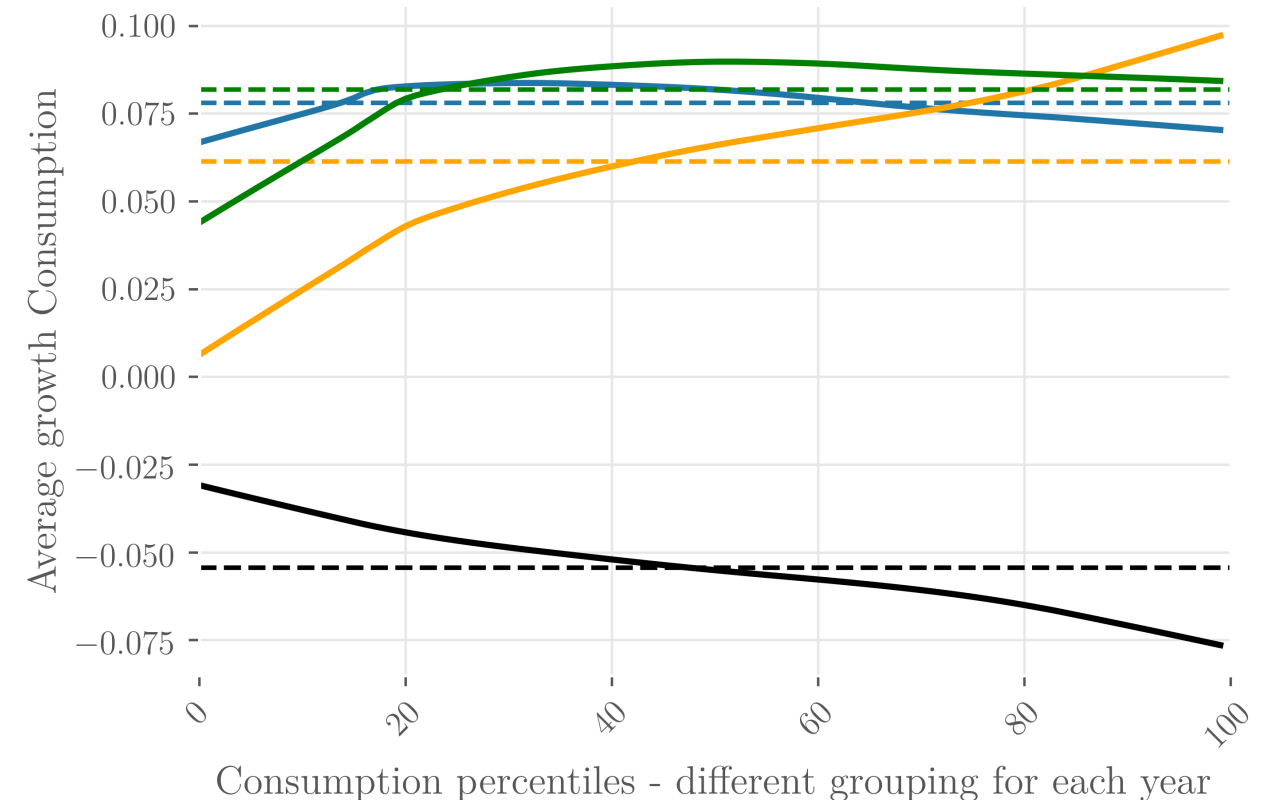
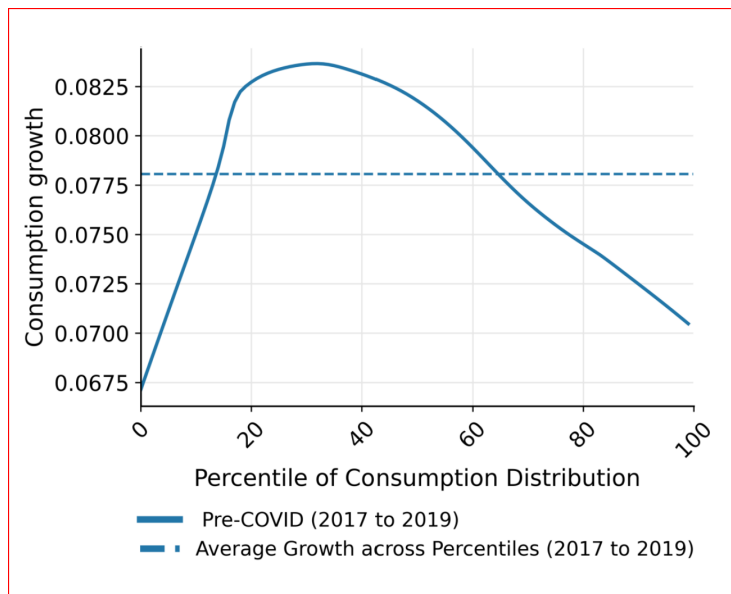
*30 days sampled randomly



One can also look at distributional aspects per categories (who consumes what), Engel curves...

Distribution of consumption growth.

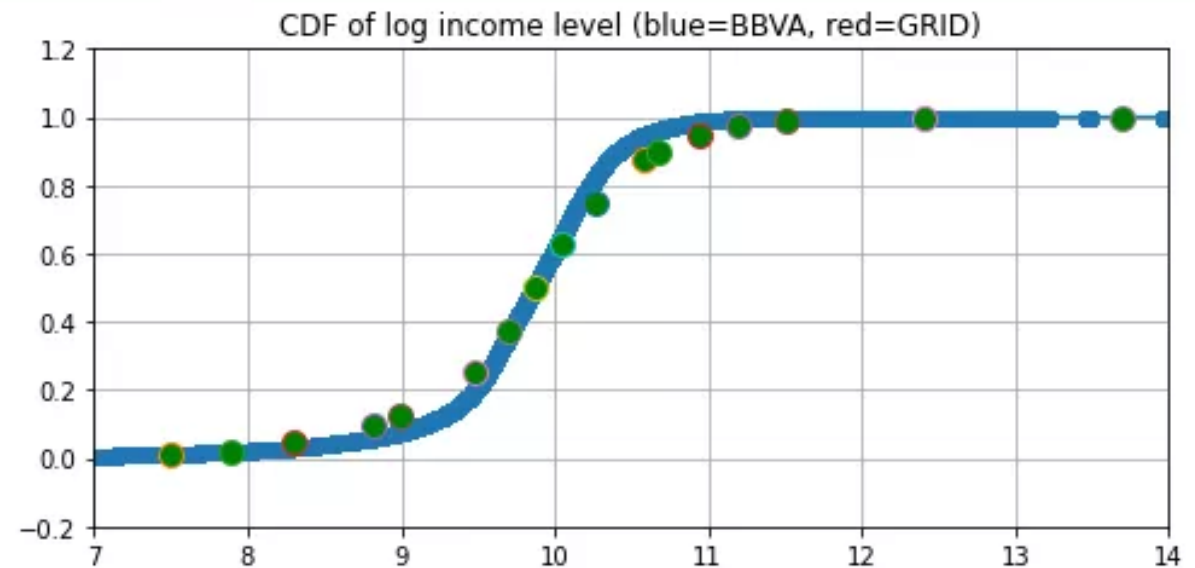
- No big change before COVID
 - Not in tails.
- No big change in overall 5 years
- But COVID roller-coaster.
- COVID: Big decrease inequality
- Recovery: Big Increase.

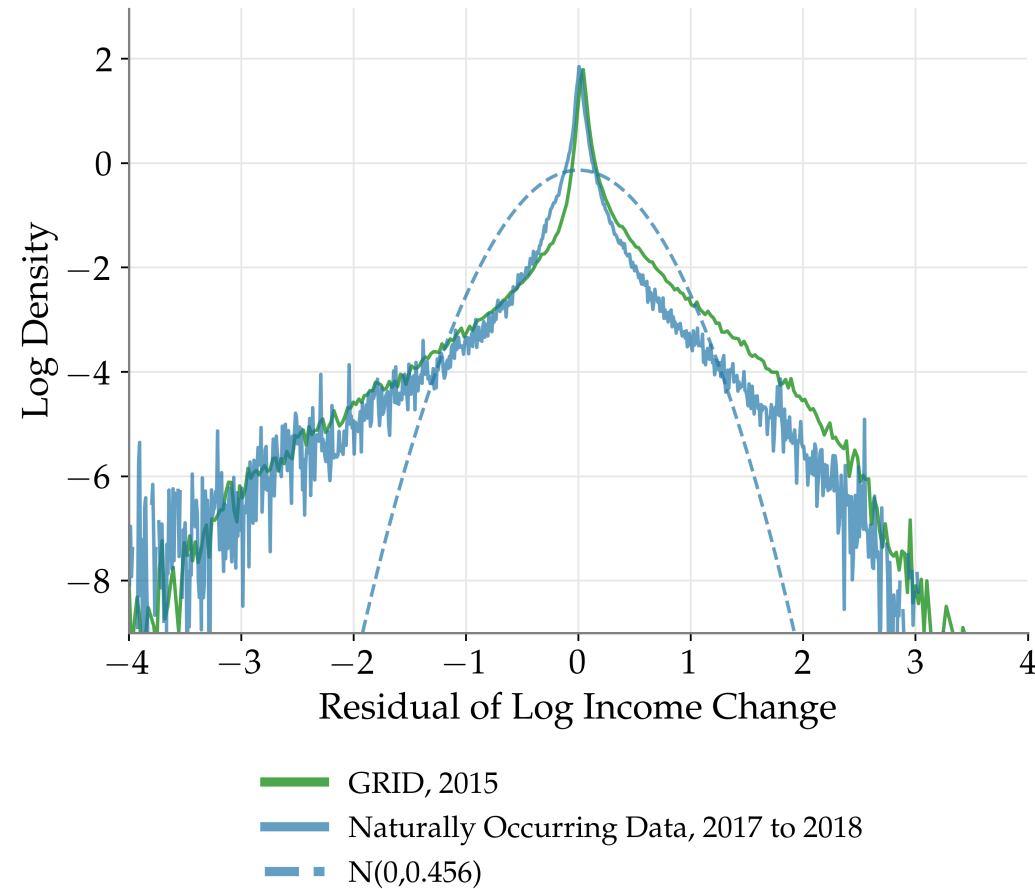


- Pre COVID 2017 to 2019
- - - 2017 to 2019 mean growth
- Post COVID 2020 to 2021
- - - 2020 to 2021 mean growth
- Post COVID 2019 to 2020
- - - 2019 to 2020 mean growth
- Whole Period 2017 to 2021
- - - 2017 to 2021 mean growth

- This paper is not about income
- ... but we have income data
- here we use it only as indirect validation
 - because we have no administrative data on consumption changes.
 - Actually, in Spain (like in many countries) there is no *good* and systematic panel of consumption.

- Not our main goal in this paper, but we can also observe income.
- Given GRID, we want to validate with their data not only on income, but on income growth.
- Our data matches GRID well
- slightly different years
- Wages and some transfers.
- Theirs is before tax, ours is kind-of-after tax





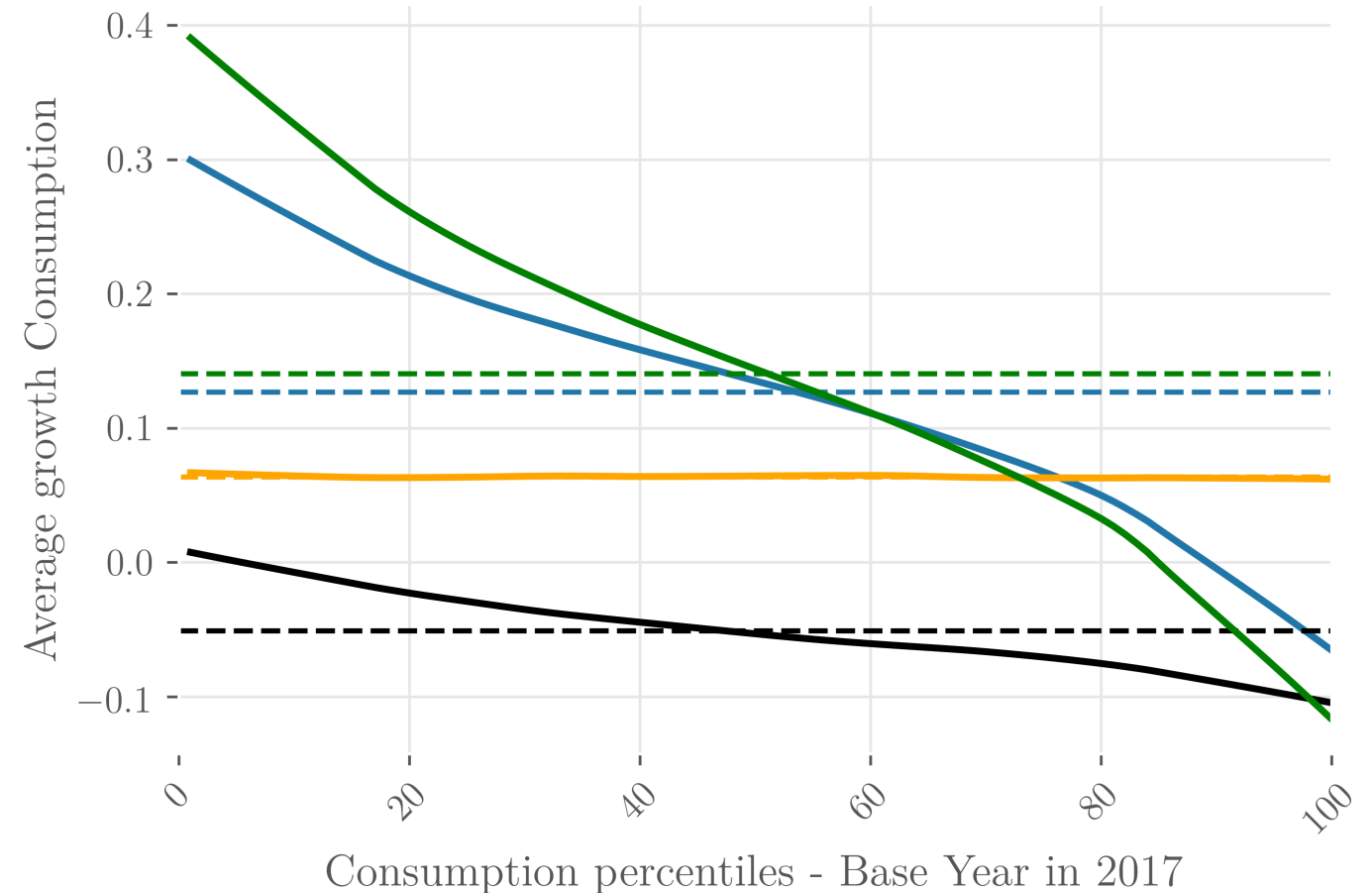
Very similar Pareto Tails:

- In left: 1.52 in our data, (lower tail power law exponent of -0.52 in the CDF); 1.58 in GRID
- In right: -2.7 (upper tail power law exponent of 1.7); -2.44 in GRID.

Dynamics of Individual Consumption

- We can exploit the individual linkages over time
- Uniquely rich data to understand consumption dynamics.
... that aggregates into national accounts

- Panel Structure at individual level.
- Very different from Cross-Sections
- In one year (pre-COVID), massive mean reversion
- 2020, 2021, much flatter... because of rapid reversion during first year.

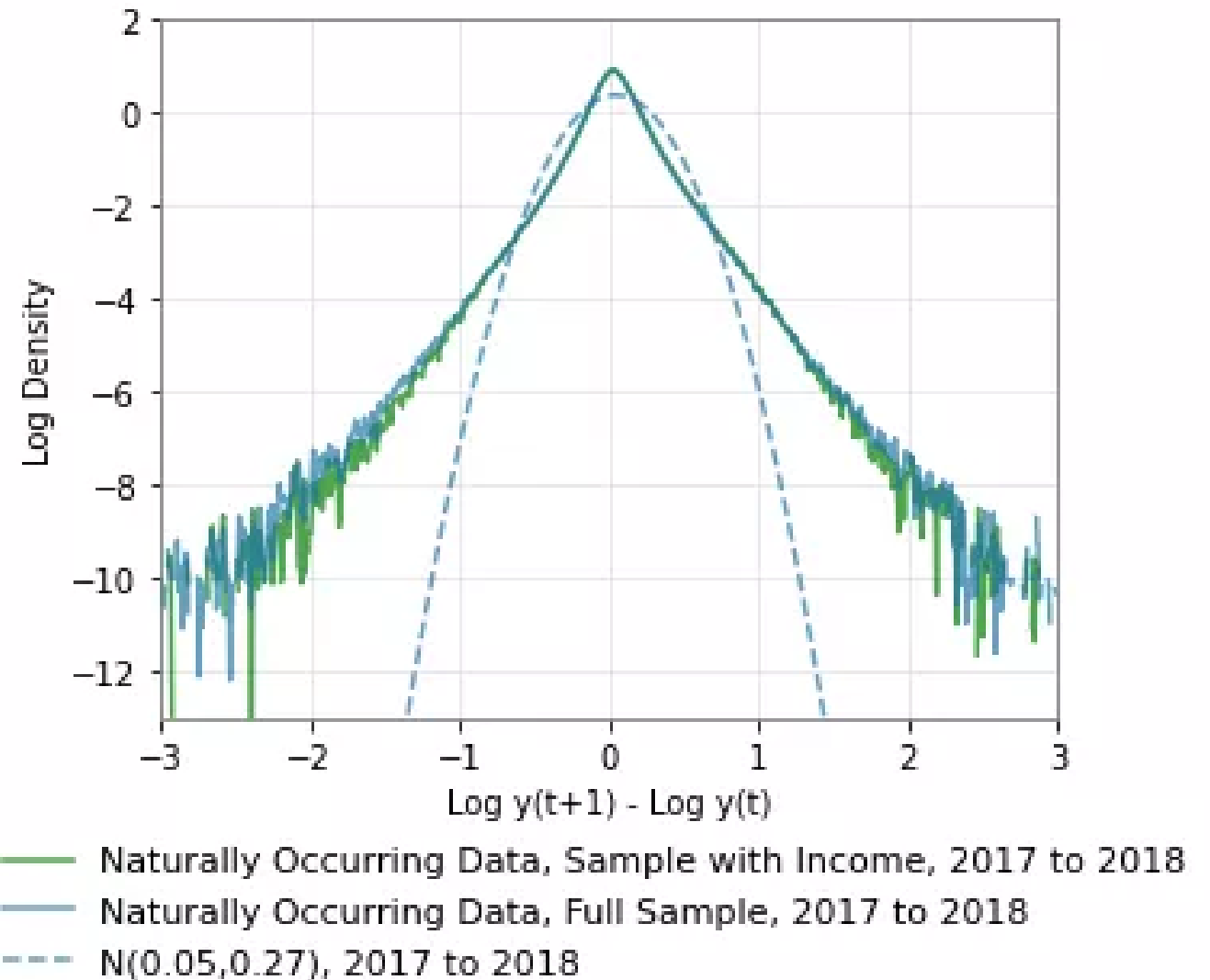


- Pre COVID 2017 to 2019
- Post COVID 2020 to 2021
- - - 2017 to 2019 mean growth
- - - 2020 to 2021 mean growth
- Post COVID 2019 to 2020
- Whole Period 2017 to 2021
- - - 2019 to 2020 mean growth
- - - 2017 to 2021 mean growth

The distribution of consumption growth does not look Gaussian

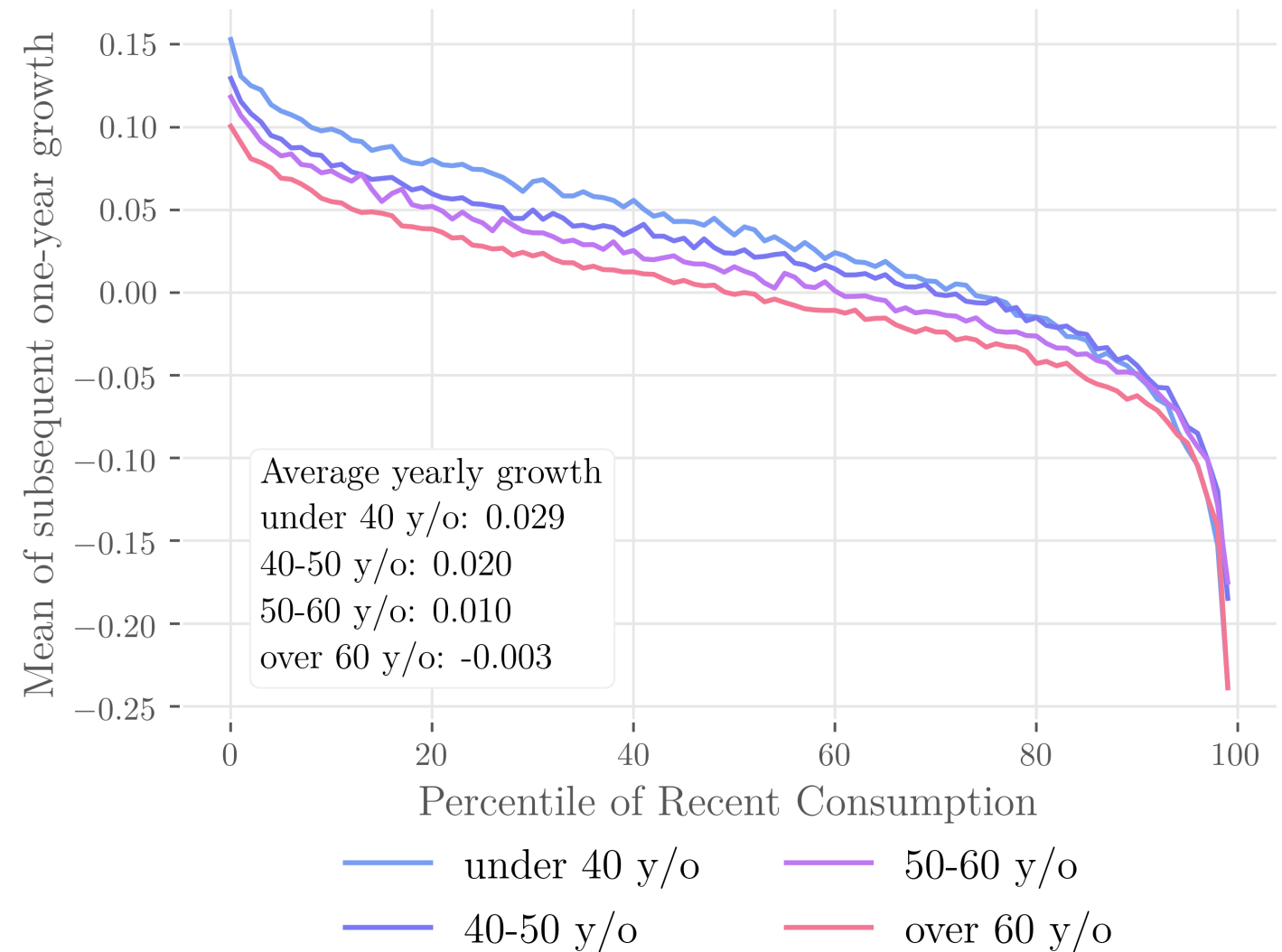
- Thick Tails
- Albeit less than income.
- Very non linear and lumpy process.
- With very strong mean reversion.
- Possible causes:
 - Income Process
 - Lumpiness of purchases themselves (frequency)

Data aggregates into national accounts.

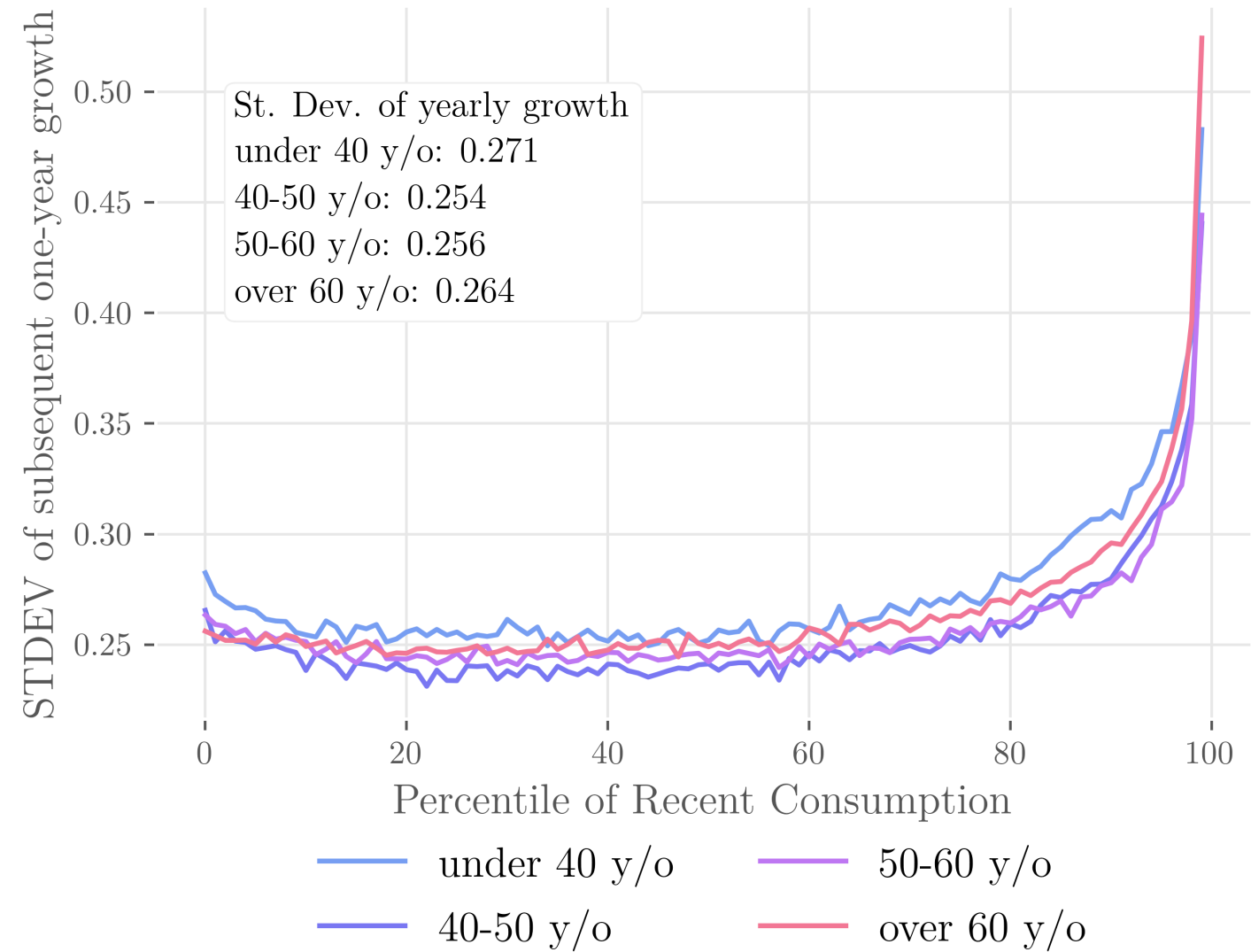


Mean-reversion of consumption growth

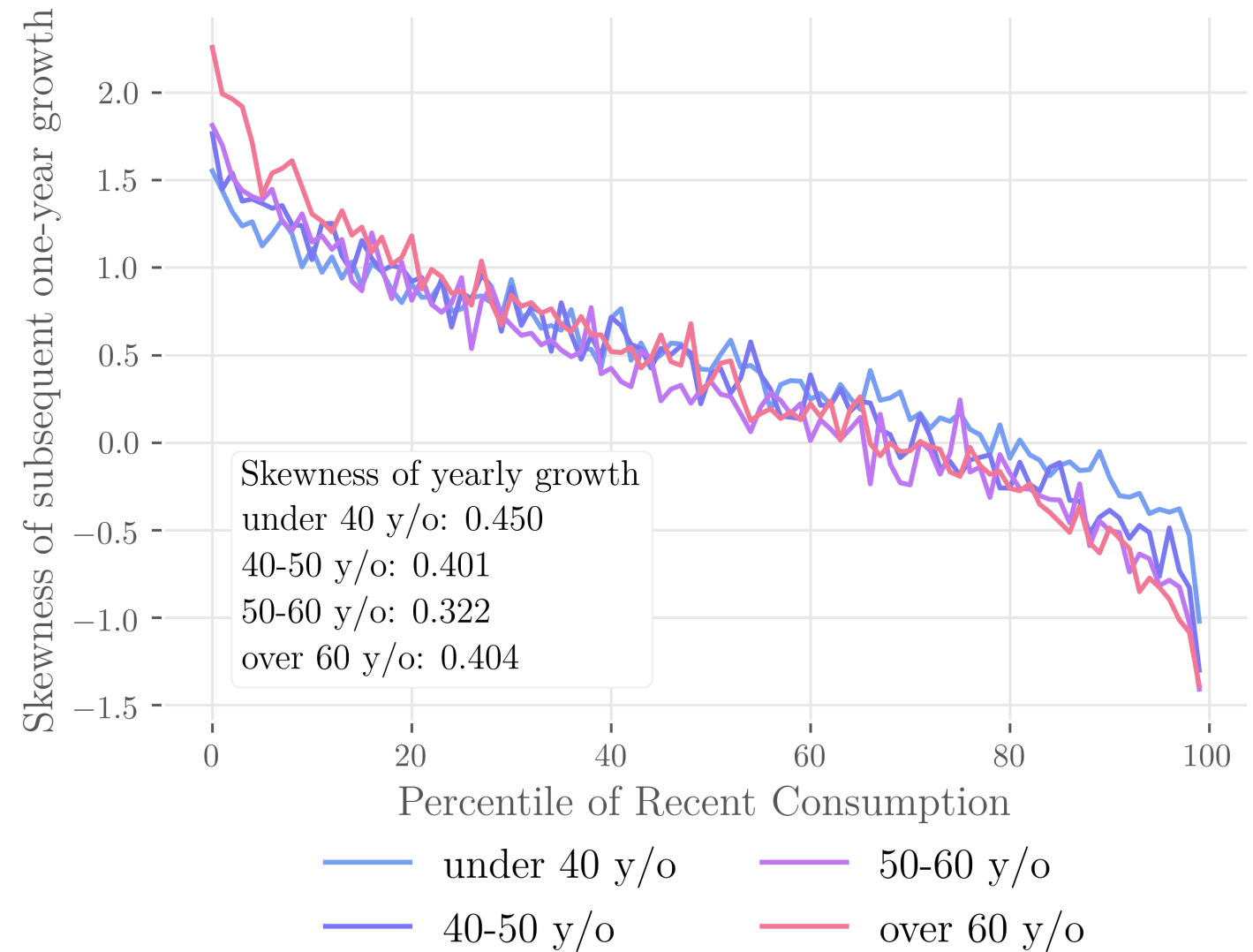
- Average growth rate per percentile of consumption.
- Young have more consumption growth



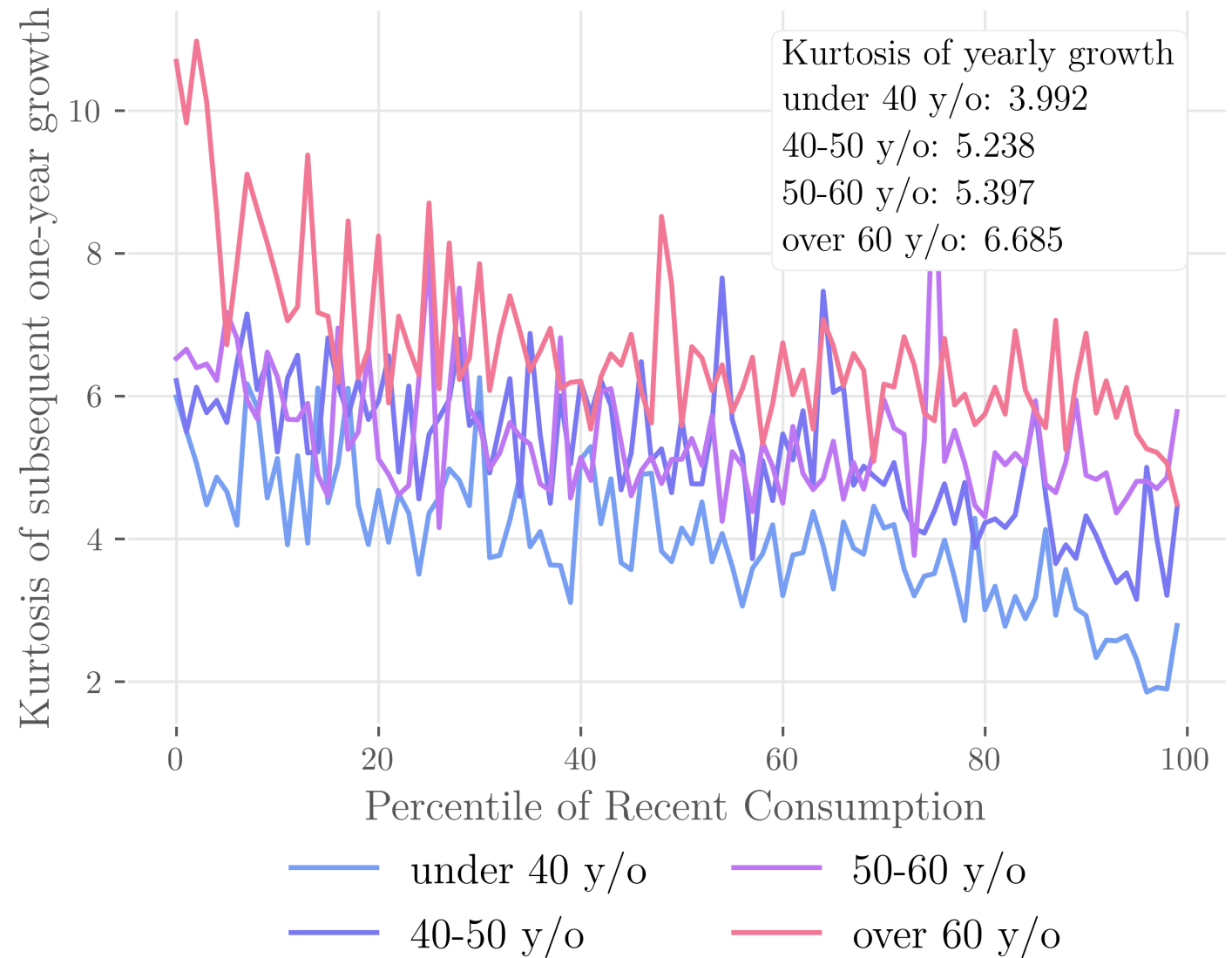
Unpredictable growth of those with very large consumption



- Left tailed distribution of growth for those in high percentiles.
 - Mostly small decreases
 - But some large huge ones
- ... Right tailed growth for those in low percentiles.
 - Mostly small increases
 - but some huge ones.

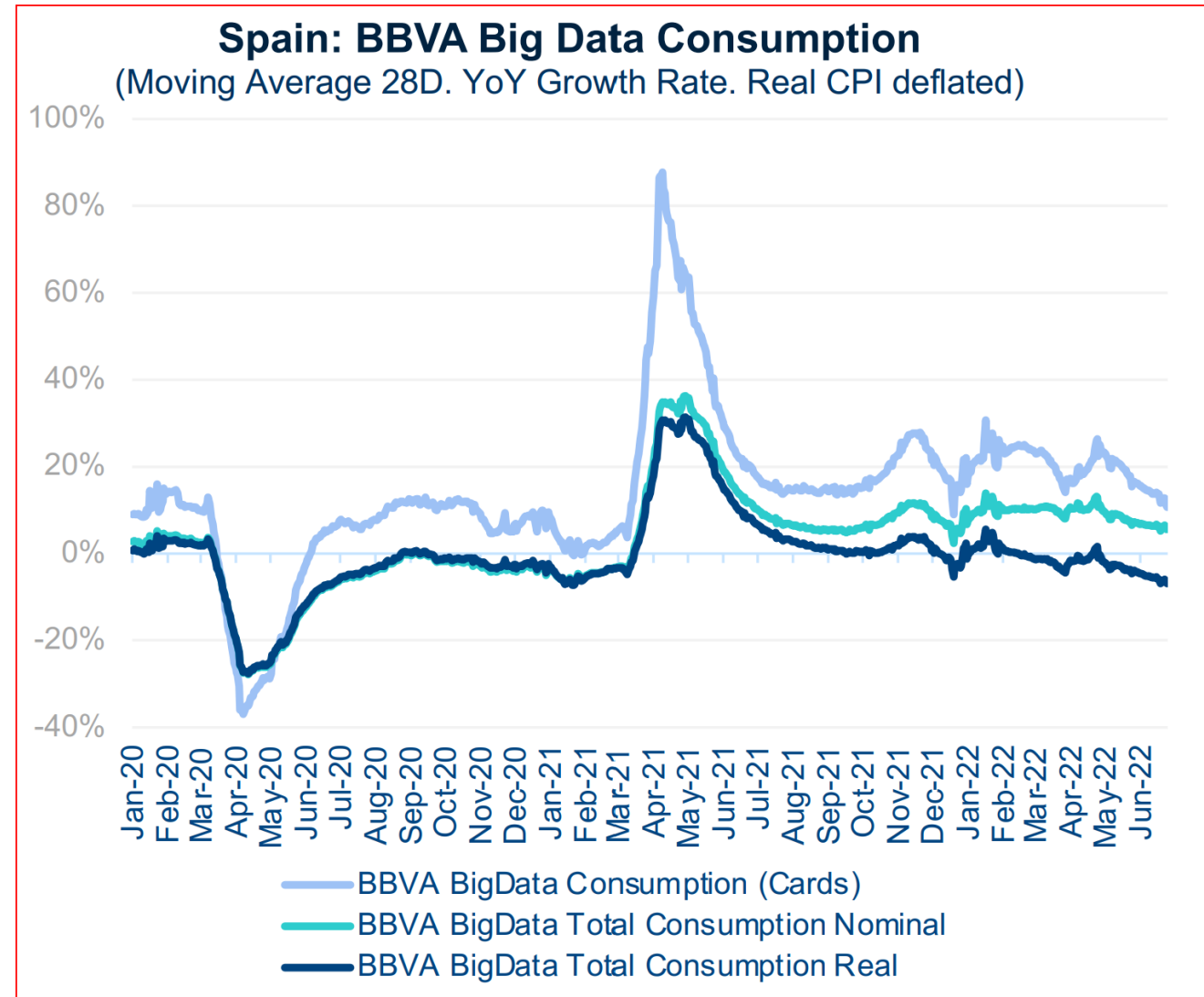


- Excess Kurtosis of consumption growth.
 - Particularly for older
 - And low consumption.



- The vast amount of data naturally occurring within financial institutions can be harnessed to produce high quality consumption survey.
- Unlike standard consumption surveys, a simple aggregation of the survey's data results in National Accounts levels
- But with arbitrary frequency, and incredibly more dense coverage.
- Not only the survey micro data generates distributional accounts for consumption
... it allows an individual panel structure that allows for careful study of consumption dynamics
- Of course, it allows for using covariates (income, ...) to understand the determination of consumption at micro and macro levels.

- In a companion paper (with added co-authors) we use daily frequency of the series to measure reaction time to monetary policy shocks.
- We ask:
 - At which frequency does monetary policy operate?
 - Does aggregation into lower frequency mask short-run effects?
- Daily aggregate consumption: 01/04/2015 to 31/12/2021.
- 90 day backward moving average.



Main Estimation Equation

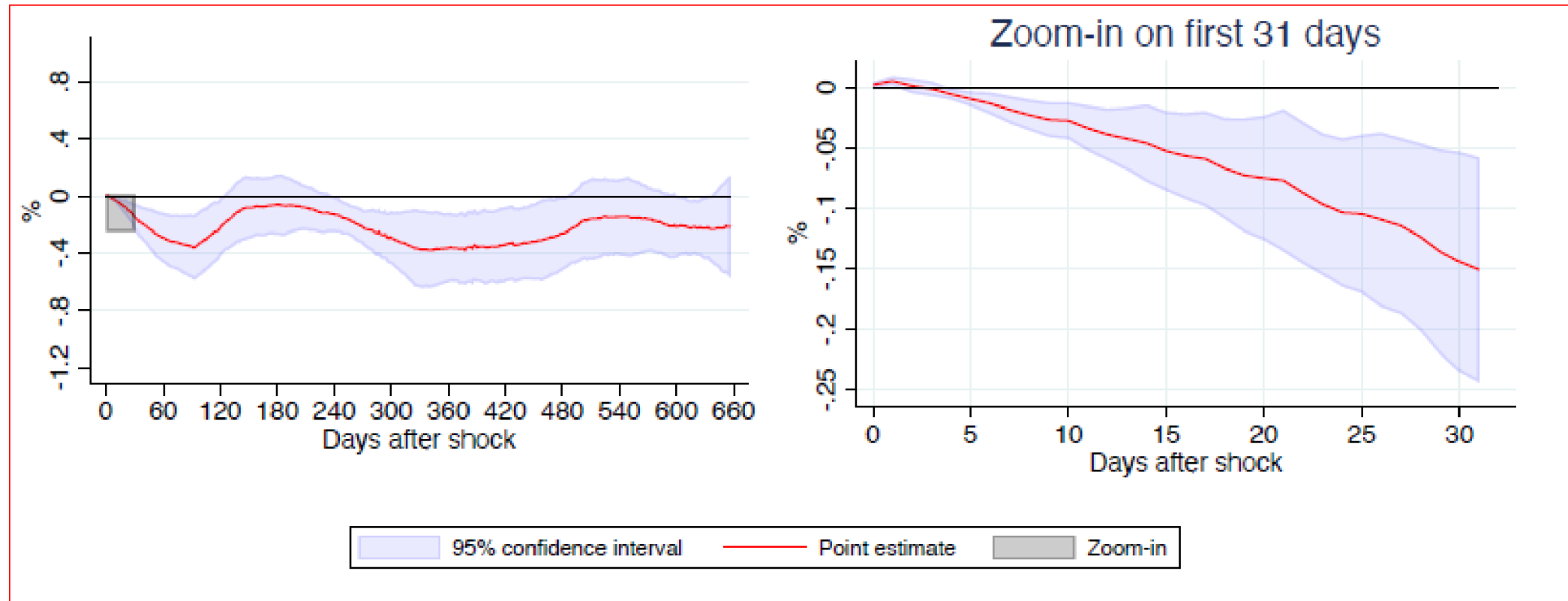
$$y_{t+h} = \alpha_h + \beta_h shock_t + \sum_{\ell=1}^p \varphi_{h,\ell} y_{t-\ell} + \theta_h cases_t + \delta_h stringency_t + \varepsilon_{h,t}$$

Variables

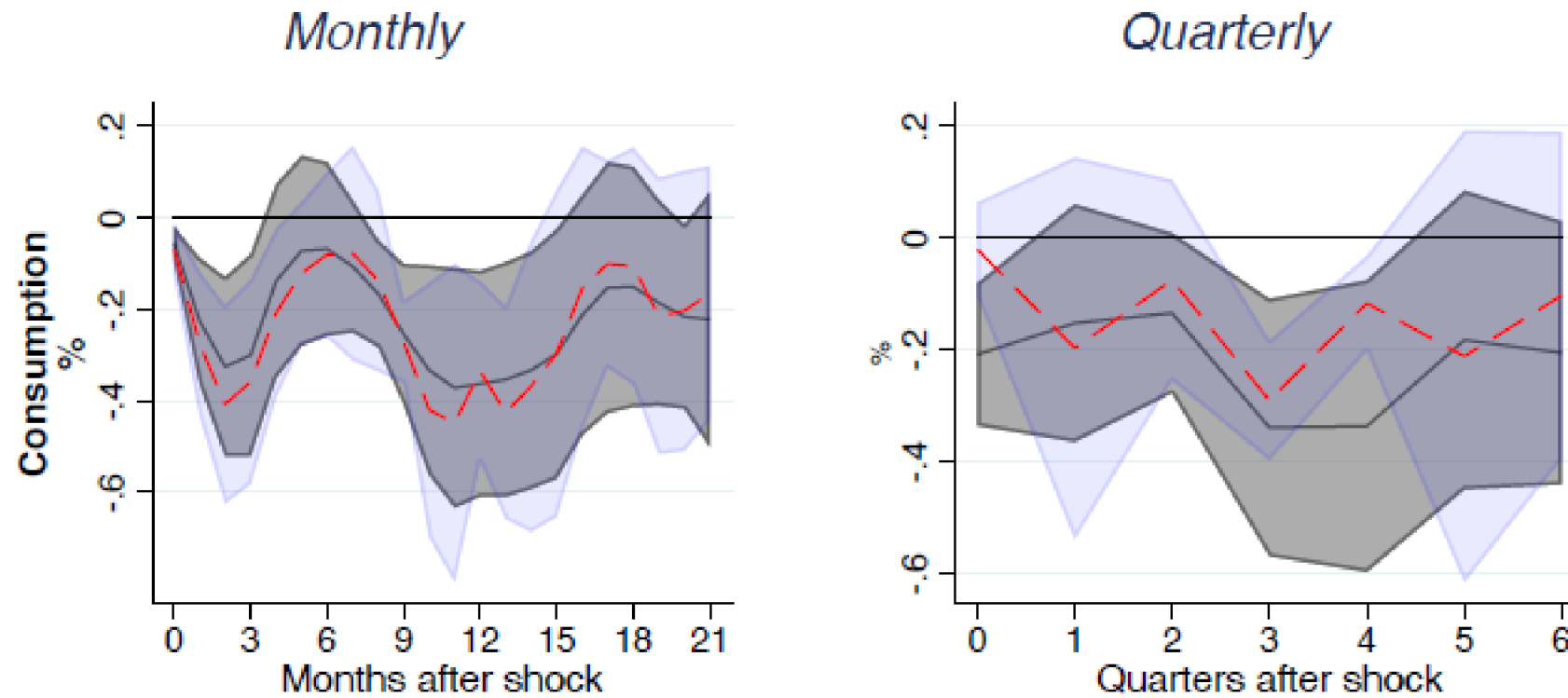
- *shock* is from Altavilla et. al. (2019) with focus on one-year yield around ECB monetary announcements
- *cases* is log daily COVID cases in Spain
- *stringency* is control for COVID-related restrictions

Model Parameters

- h varies from 1 to 658
- $p = 90$



Consequences of Time Aggregation



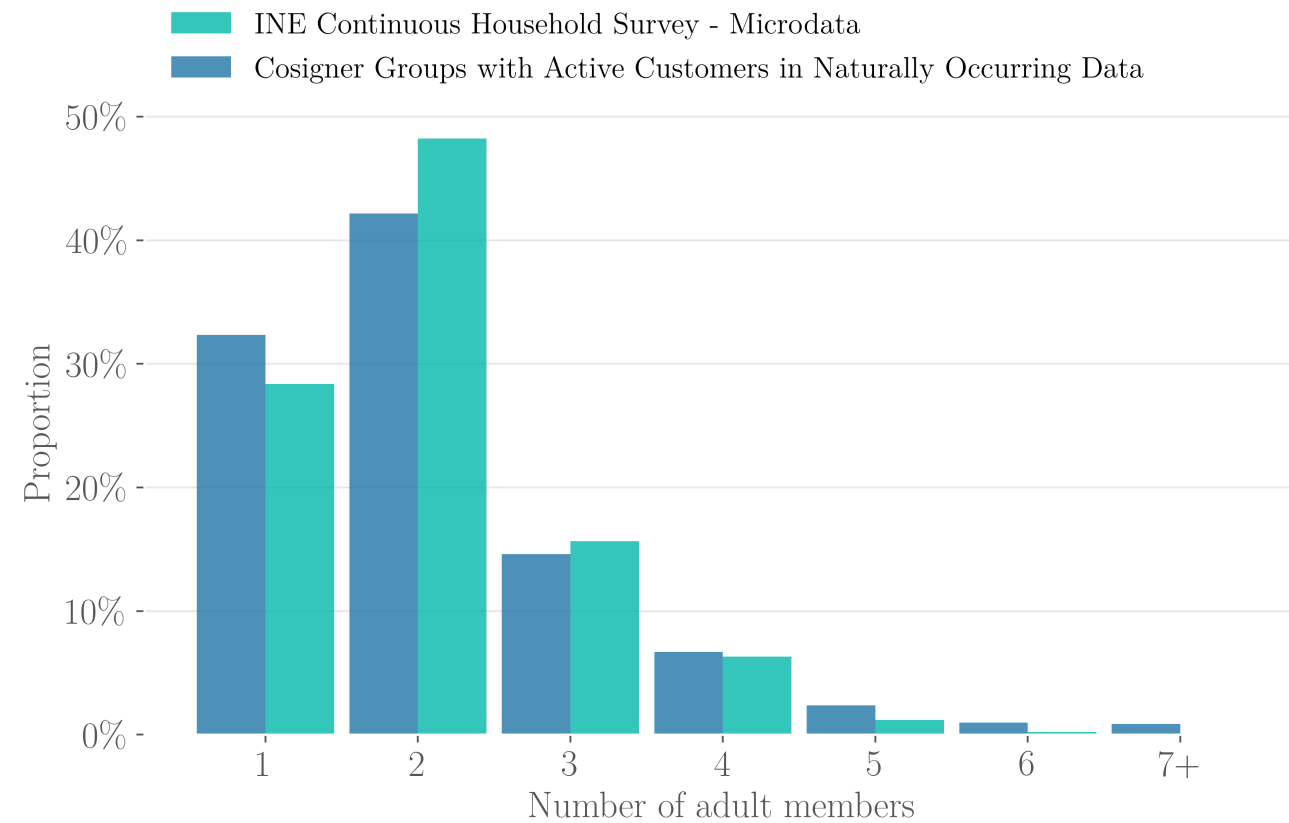
Solid-grey IRF: daily IRF aggregated to lower frequency

Red-dash IRF : source data aggregated to lower frequency and then IRF

Christiano-Eichenbaum (1987): “temporal aggregation bias can be quantitatively important in the sense of significantly distorting inference”

Extra Material

- The bank does not provide household linkages or specific addresses due to legal issues
- We construct household relationships
 - Network of people who share accounts and lives in the same census tract.
 - We know marriage status. If unlinked, we add one person.
 - We know the number of dependent adults, if unfilled, we add up to that number.
- Our distribution seems reasonable.



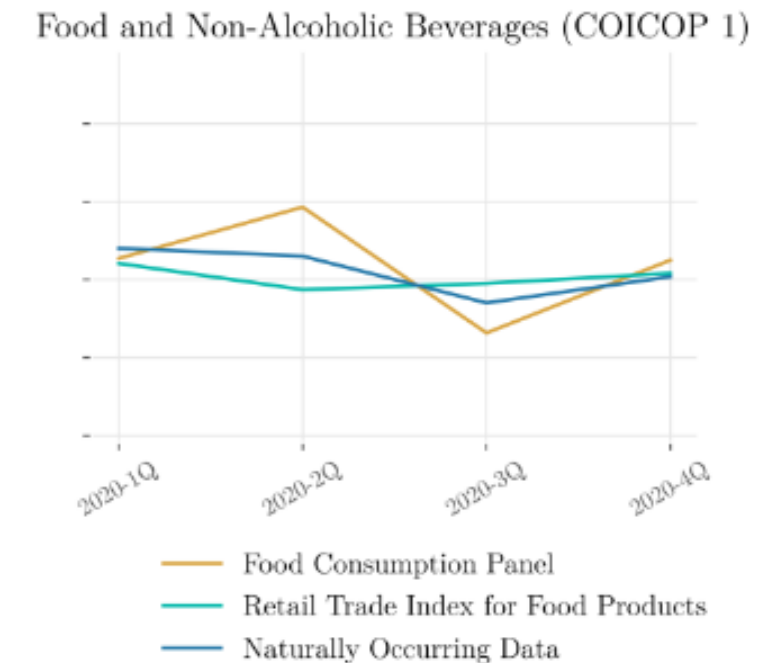
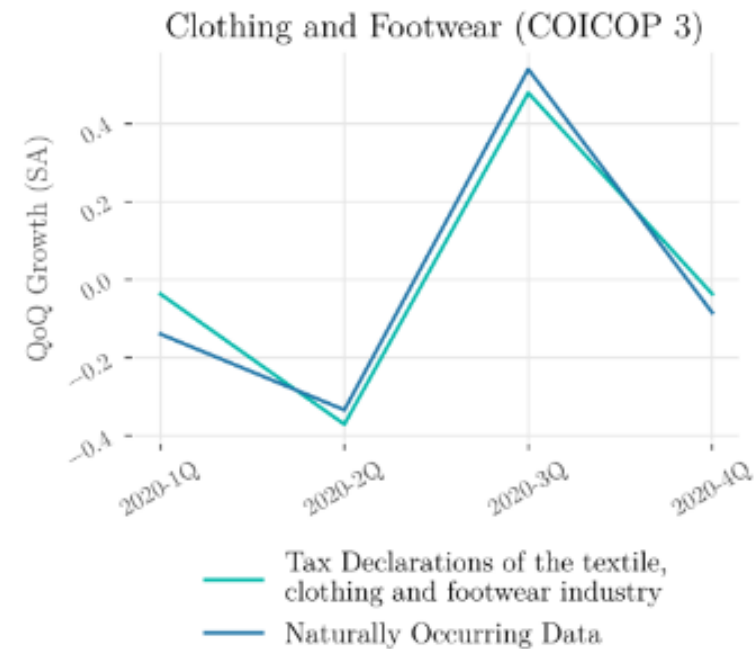
- We observe individual spending, not consumption, within households
- We observe household **housing** spending
- ASSUME equal spending among active clients within households and half the weight of non-active clients.

$$c_i = \frac{\sum_{j \in A(i)} c_j^{\text{NH}} + c_{h(i)}^{\text{H}}}{A(i) + 0.5O(i)}.$$

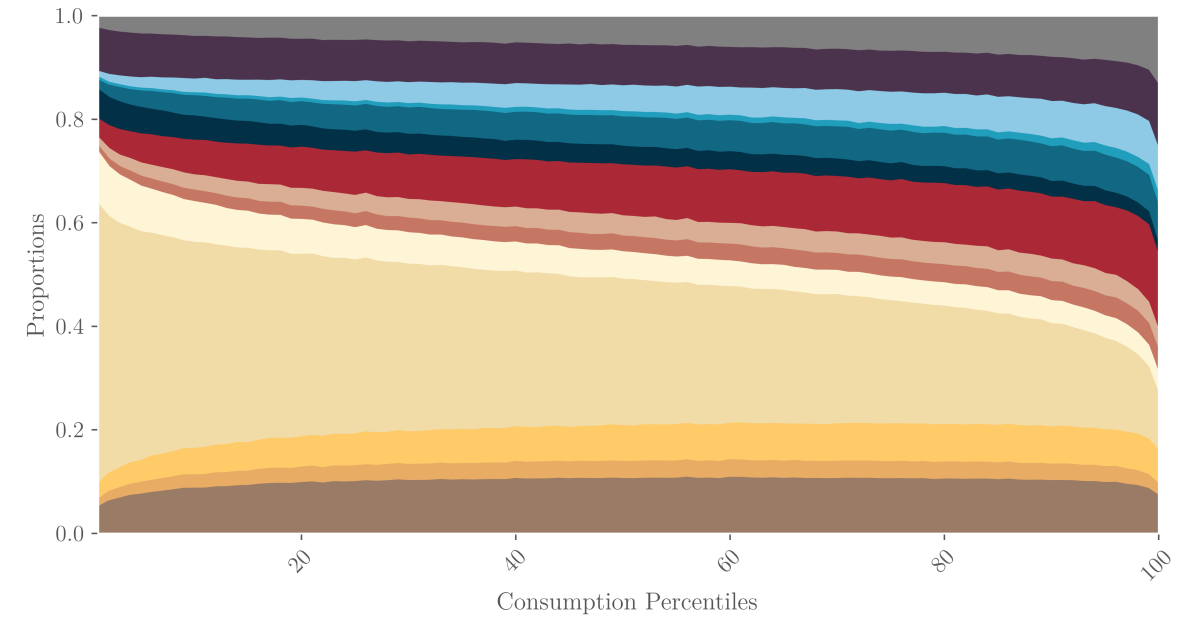
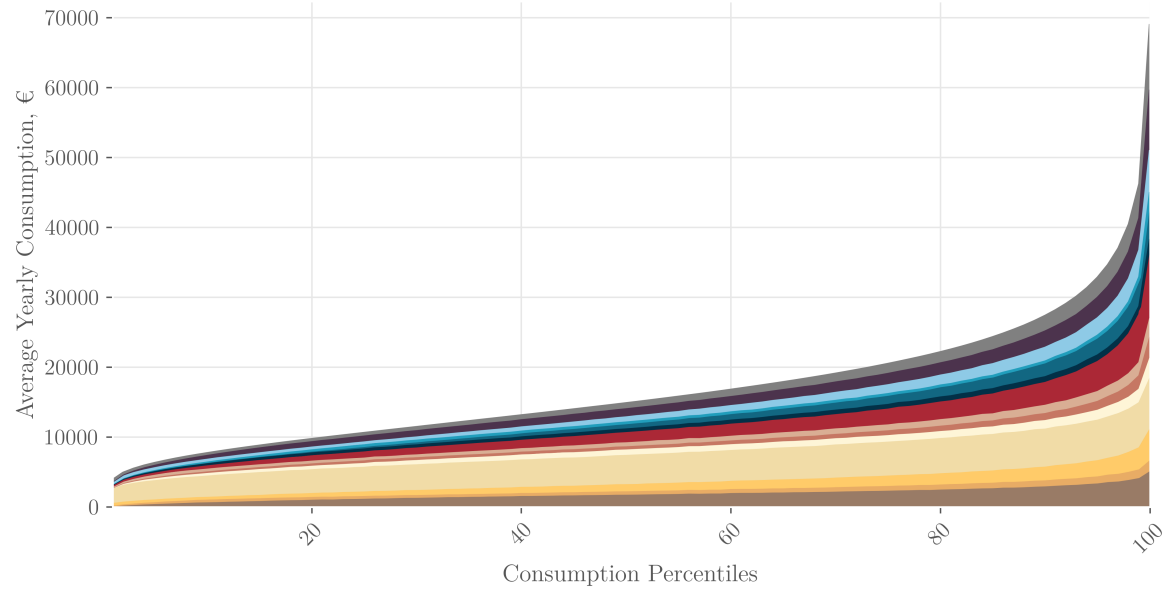
	National Accounts	Baseline	Weight of non-actives = 0	Weight of non-actives = 0.25	Weight of non-actives = 0.75	Weight of non-actives = 1	No Demographic Weights Applied	Dropping Imputed Rent	Dropping Online Card Transactions	Dropping Cash Withdrawal	Only Card and Cash Withdrawal with Card
Levels	Correlation with Nat. Accs.	0.738	0.792	0.760	0.720	0.708		0.798	0.870	0.443	0.624
	Rooted MSE vs. Nat. Accs.	8040	47013	18719	14270	21743		42988	10052	39645	92567
	Mean (million €)	164357	164118	210636	181235	152560	144154	121857	155546	127079	72196
	Standard Deviation (million €)	9706	11832	13453	12445	11368	11064	10501	9111	14613	10252
QoQ	Correlation with Nat. Accs.	0.968	0.969	0.968	0.968	0.968	0.964	0.962	0.961	0.939	0.906
	Rooted MSE vs. Nat. Accs.	0.025	0.025	0.025	0.025	0.025	0.025	0.019	0.023	0.037	0.034
	Mean	0.007	0.011	0.009	0.010	0.011	0.012	0.009	0.014	0.008	0.017
	Standard Deviation	0.064	0.043	0.042	0.042	0.043	0.043	0.060	0.046	0.032	0.069

Table C.3: Impact of Modeling Choices on Relationship between Official and Naturally Occurring Consumption Series

- Different measurement
- We just sum data.
- INE uses some surveys and some administrative data.
- We diverge in COPICOPS where they use surveys plus model.



Distribution across COICOPS and people (1/3)



- | | |
|--|--|
| 01 Food and Non-Alcoholic Beverages | 07 Transport |
| 02 Alcoholic Beverages and Tobacco | 08 Communication |
| 03 Clothing and Footwear | 09 Recreation and Culture |
| 04 A Housing | 10 Education |
| 04 B Water, Electricity, Gas, and Other Fuels | 11 Restaurants and Hotels |
| 05 Furnishings, Household Equipment and Maintenance | 12 Miscellaneous Goods and Services |
| 06 Health | Uncategorized |

- | | |
|--|--|
| 01 Food and Non-Alcoholic Beverages | 07 Transport |
| 02 Alcoholic Beverages and Tobacco | 08 Communication |
| 03 Clothing and Footwear | 09 Recreation and Culture |
| 04 A Housing | 10 Education |
| 04 B Water, Electricity, Gas, and Other Fuels | 11 Restaurants and Hotels |
| 05 Furnishings, Household Equipment and Maintenance | 12 Miscellaneous Goods and Services |
| 06 Health | Uncategorized |



Percentile	0.1	0.5	0.9	0.999
Necessities	0.67	0.57	0.49	0.29
Luxury	0.33	0.43	0.51	0.71

Necessities

Food and Non-Alcoholic Beverages (01); Alcohol and Tobacco (02); Clothing and Footwear (03); Housing and Utilities spending (04); and Health (06)

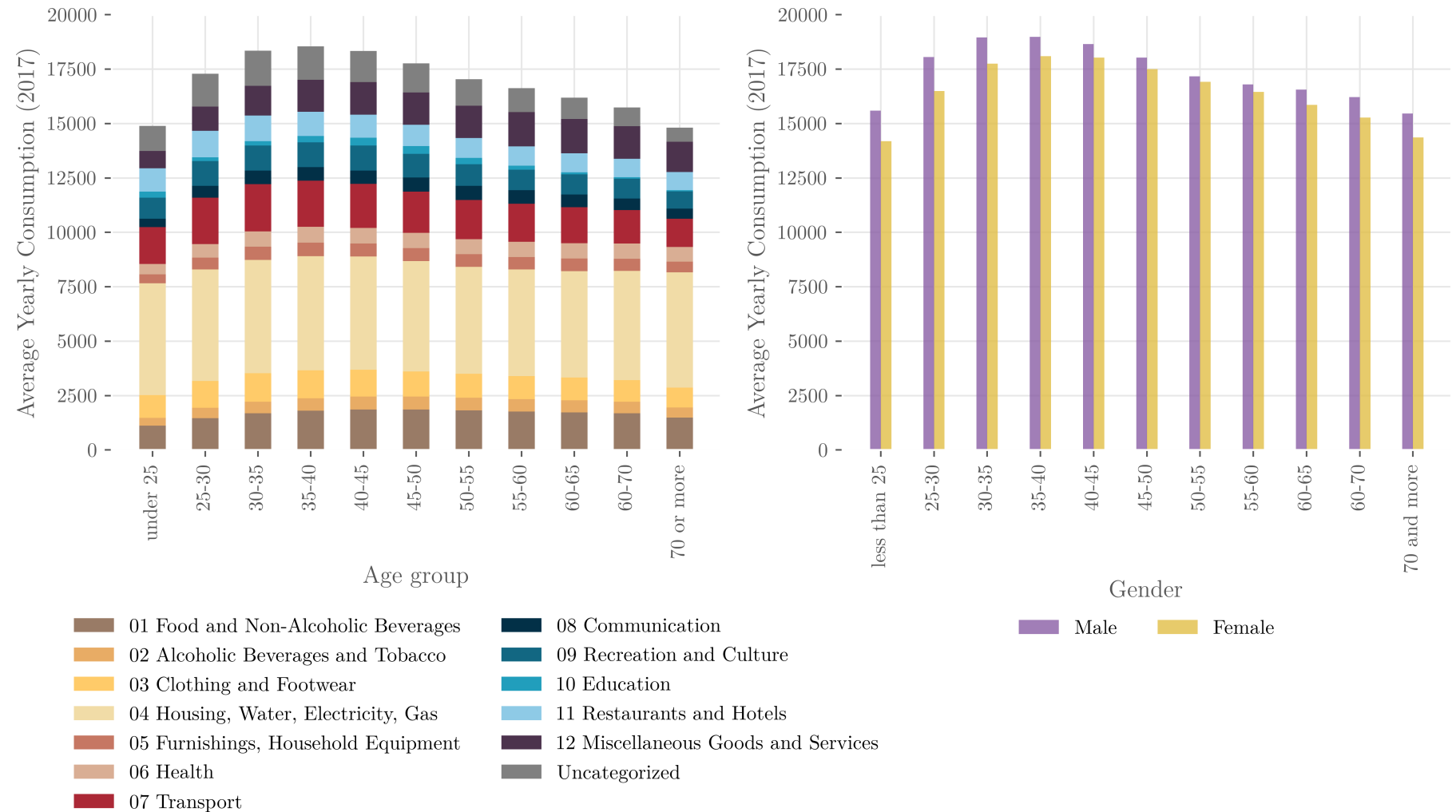
Luxuries

Furnishings and Household Equipment (05); Transport (07); Communication (08); Recreation and Culture (09); Education (10); Restaurants and Hotels (11); Miscellaneous Goods and Services (12); Uncategorized

Distribution of consumption across age and gender.

- Age profile consistent with US (Aguiar & Hurst (2013))

- 6% gender consumption gap
 - in spite of household division
 - Smaller at middle age



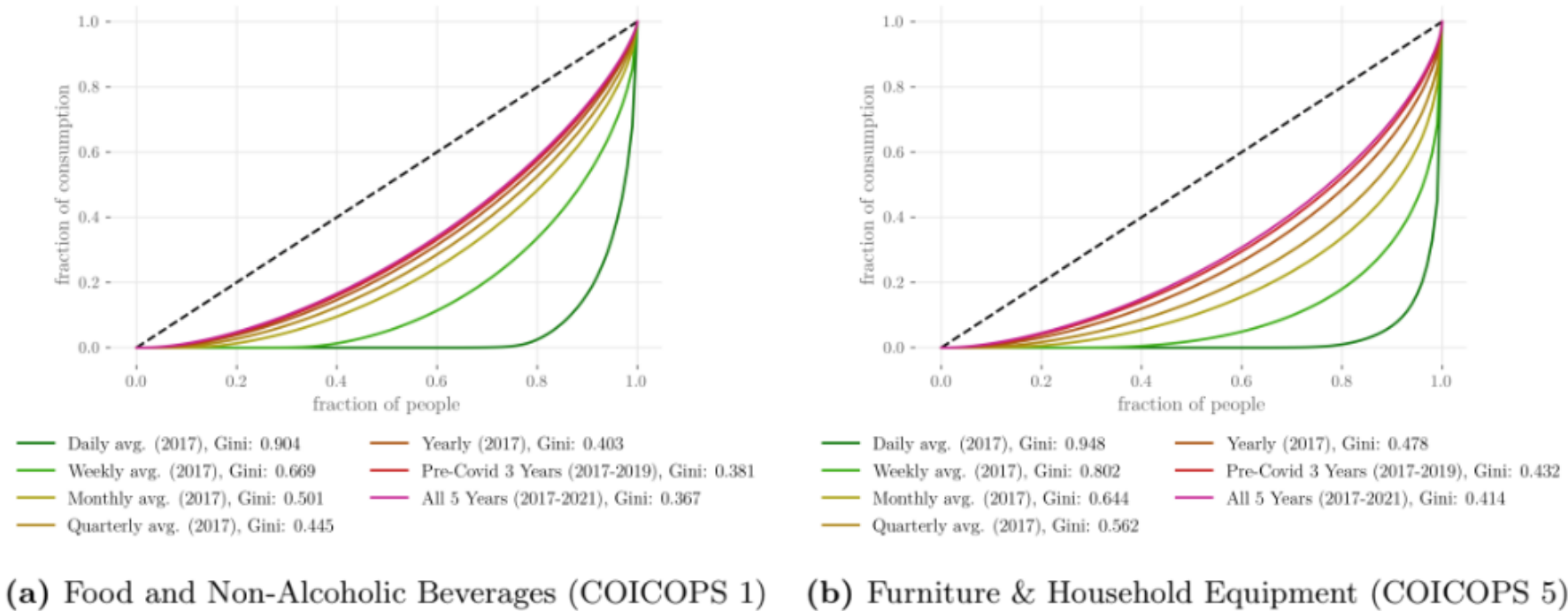
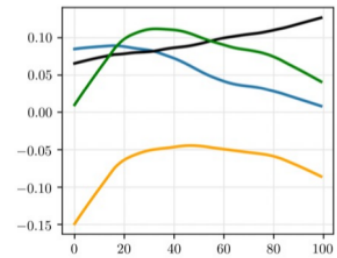
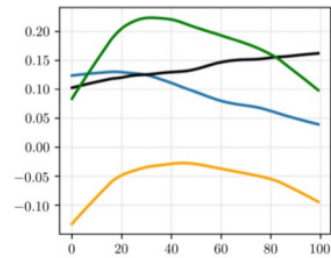


Figure C.6: Lorenz Curves and Gini Coefficients implied by the distribution of consumption of selected COICOP categories across time frequencies. Panel (a): Food and Non-Alcoholic Beverages. Panel (b) Furniture and Household Equipment.

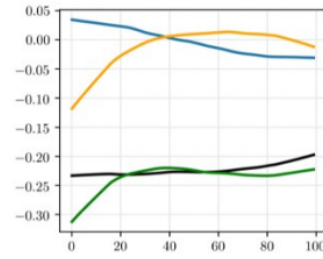
Growth Rate of Consumption per Percentiles by COICOP Consumption Group



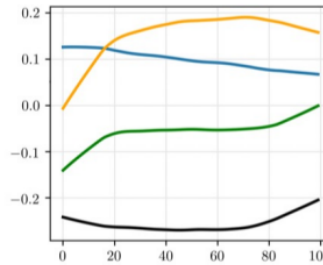
(a) Food and Non-Alcoholic Beverages



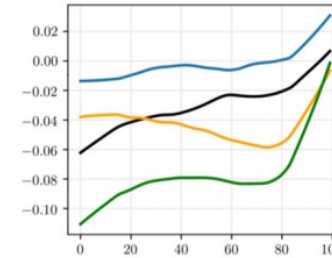
(b) Alcoholic Beverages, Tobacco, and Narcotics



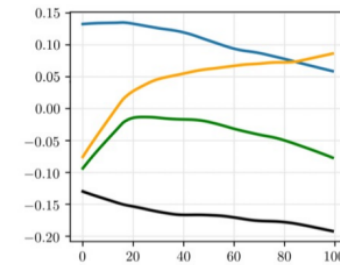
(c) Clothing and Footwear



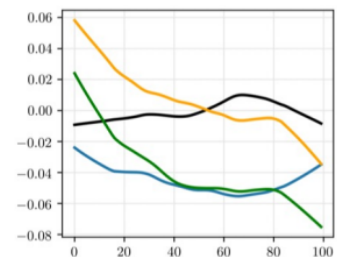
(g) Transport



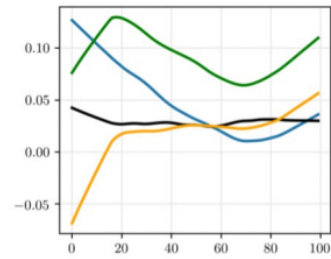
(h) Communication



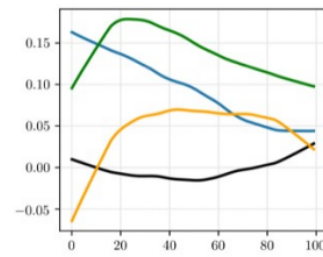
(i) Recreation and Culture



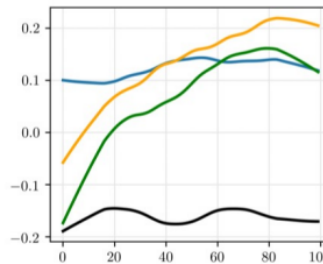
(d) Utilities (excluding imputed housing rent)



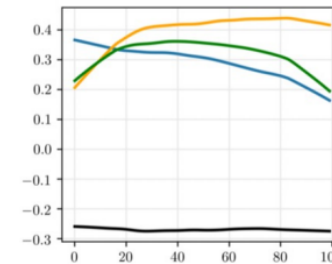
(e) Furnishings, Household Equipment, and Routine Household Maintenance



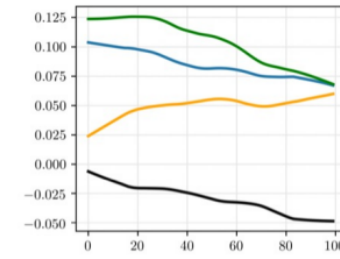
(f) Health



(j) Education



(k) Restaurants and Hotels

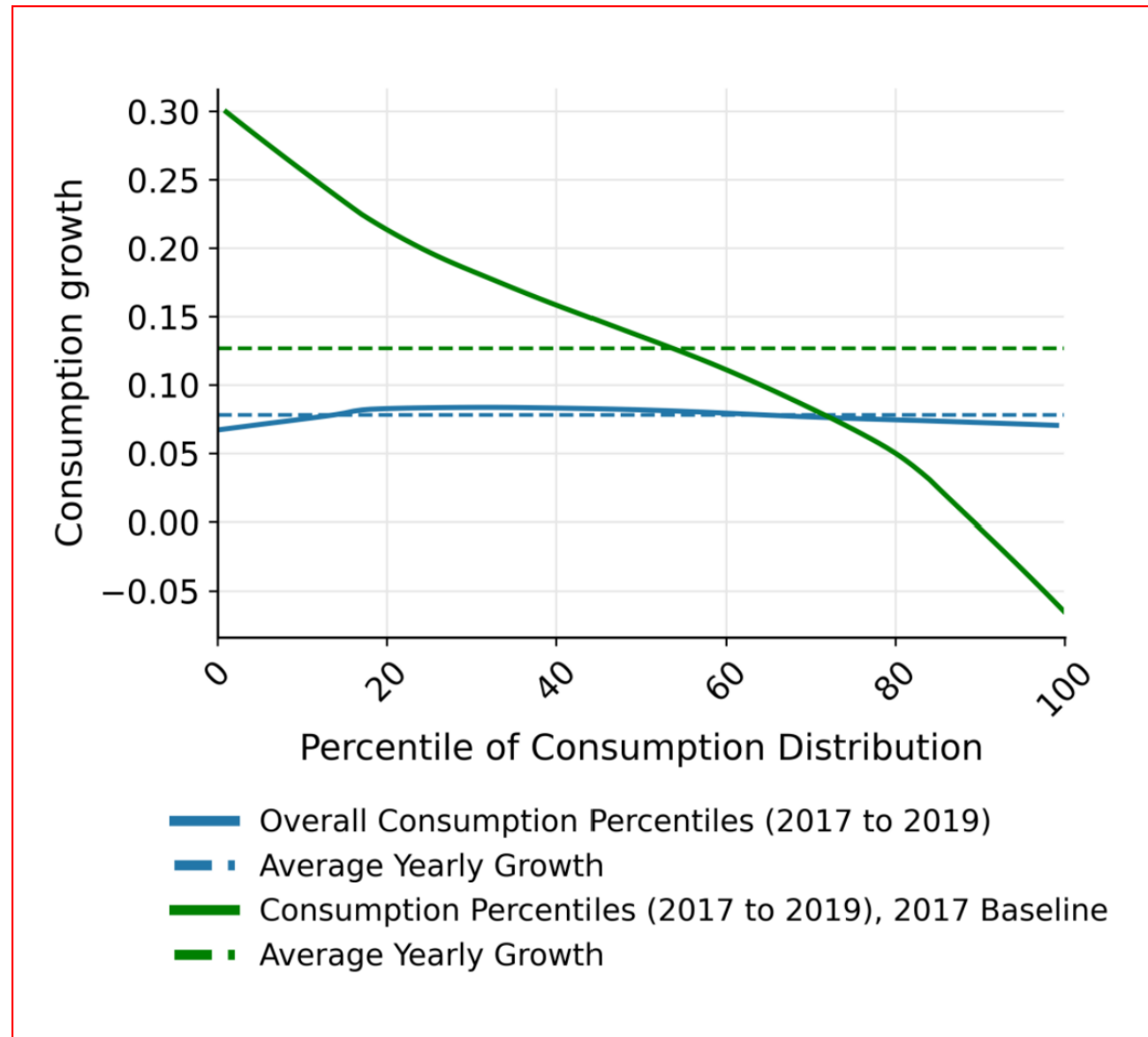


(l) Miscellaneous Goods and Services

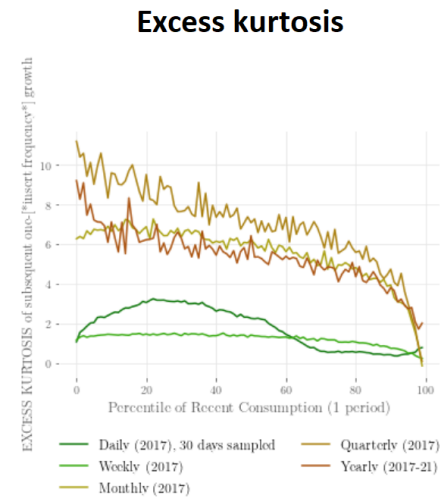
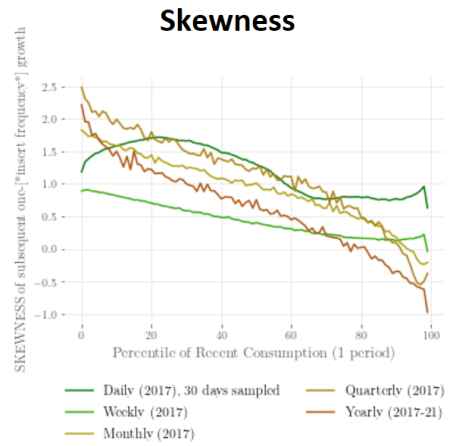
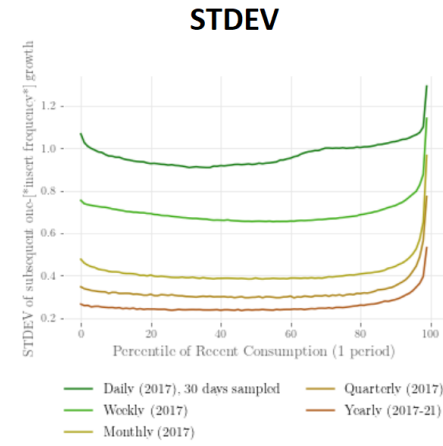
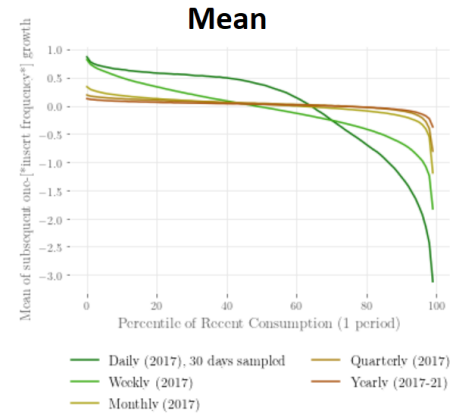
— Pre COVID 2017 to 2019
 — Post COVID 2019 to 2020

— Post COVID 2020 to 2021
 — Whole Period 2017 to 2021

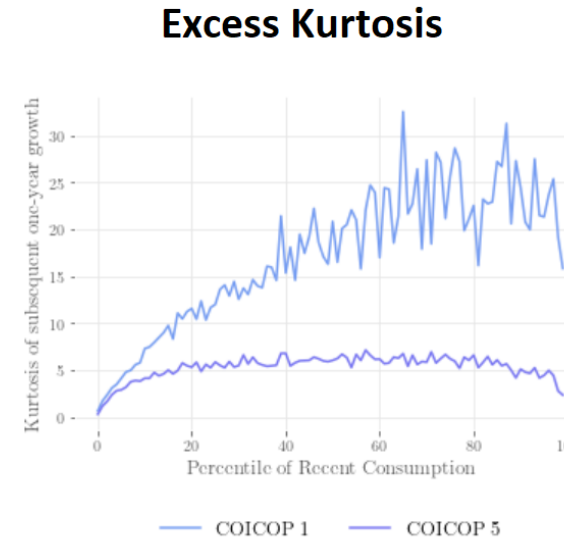
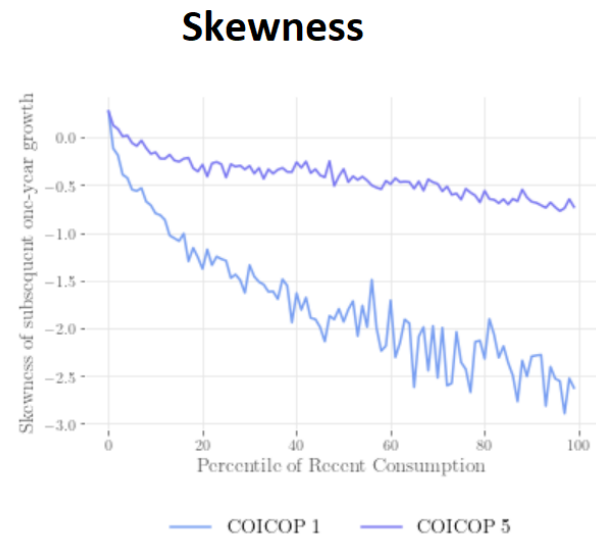
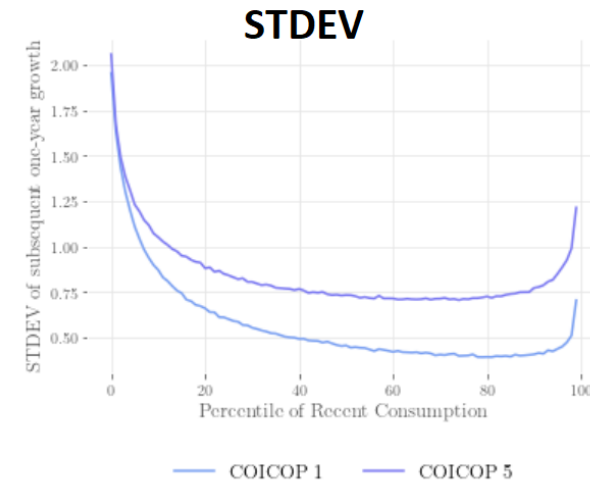
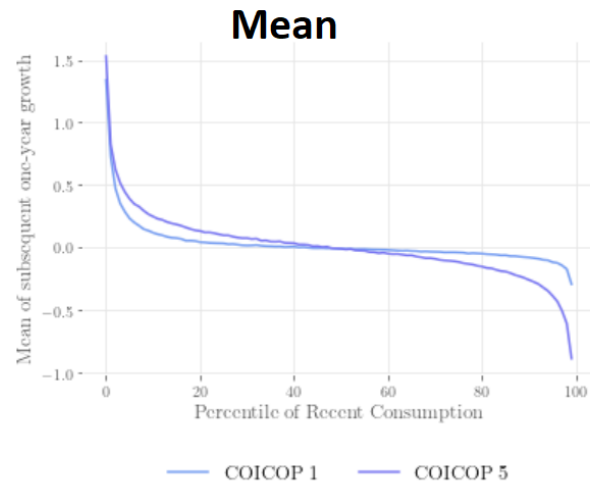
Two perspectives of Growth and Inequality



Frequencies



COICOPs



National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption

Gergely Buda (BSE)
Vasco M. Carvalho (Cambridge)
Stephen Hansen (UCL)
Álvaro Ortiz (BBVA Research)
Tomaso Rodrigo (BBVA Research)
José V. Rodríguez Mora (Edinburgh)

July 19, 2023
