High frequency tax information for tracking retail: challenges and applications

Joaquín Pérez
Statistics Division

IV Statistics Conference: “Post-pandemic Statistics”, Central Bank of Chile. September 28th 2021.

I. Context

The covid-19 crisis increased the demand for high frequency statistics.

Demand for high frequency statistics

- Decision-makers face a gap in their information set.
- New breakdowns are needed to evaluate impact of shocks.

New data sources available

- Transactional data (POS, Scanner), Bigdata (webscrapping, GPS).
- Administrative records (electronic invoices & receipts).

Challenges for Statistical Offices

International experience in the elaboration of high frequency indicators amid the covid-19 crisis.

Institution	Indicator	Data Source
INE Spain	Measurement of Large Companies Daily Retail Trade	Administrative Records
Bureau of Economic Analysis - US	Daily Spending by Industry	Credit/Debit Card Transactions
Statistics Denmark	COVID-19 Consumption Indicator (by Industry)	Credit/Debit Card Transactions

In Chile, two institutions provide short-term statistics on retail trade.

Institution	Indicator	Type	Data Source	Frecuency	Release (t+)
National Statistics Institute (INE)	Trade Activity Index - IAC (Retail)	Real	Business Surveys & Administrative Records	Monthly	29 days
National Trade Chamber (CNC)	Trade Sales Index	Nominal	Business Surveys	Monthly	29 days
	Weekly Sales Thermometer (Offline)			Weekly	7 days

II. Methodology

Project scope and data source.

Goal

- To provide with a **fast, daily index of retail sales** for the Chilean economy that **complements** the official monthly statistics (elaborated by INE).
    - **Daily Retail Sales Index** – *Índice de Ventas Diarias del Comercio Minorista (**IVDCM**)*

Dataset

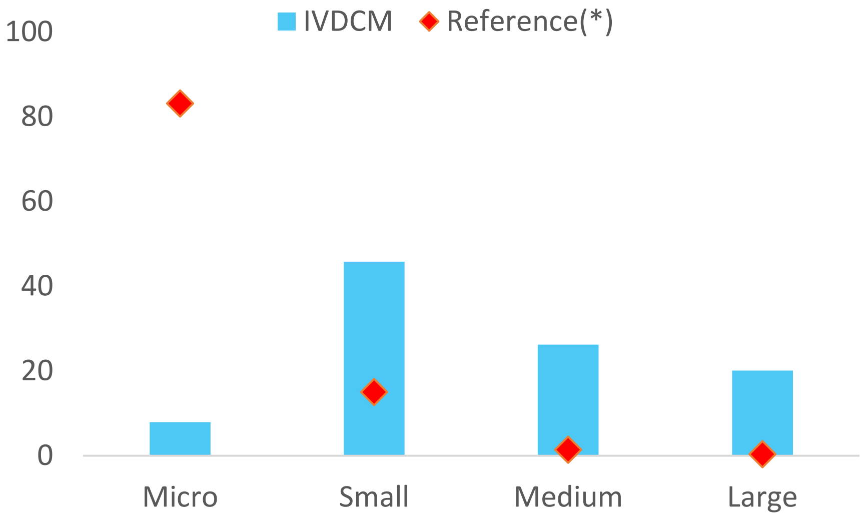
- The data comes from the Chilean Tax Agency (SII) and consists of **daily sales** of Chilean firms **authorized to emit electronic receipts**.
- Electronic receipts include **online & offline**, **card & cash** transactions between **businesses and final consumers** (B2C).
- The data is **updated weekly** with a lag of 5 days.

The sample of companies is dominated by large-sized firms, which impacts the weights by industry.

- We use the National Accounts **business register** to select retail trade companies.
- To ensure stability across time, we selected companies with **monthly presence** in the database between jan. 2018 & sept. 2019 resulting in **2,119 firms** accounting for **25.2% of total retail sales**.

Retail trade: distribution by firm size

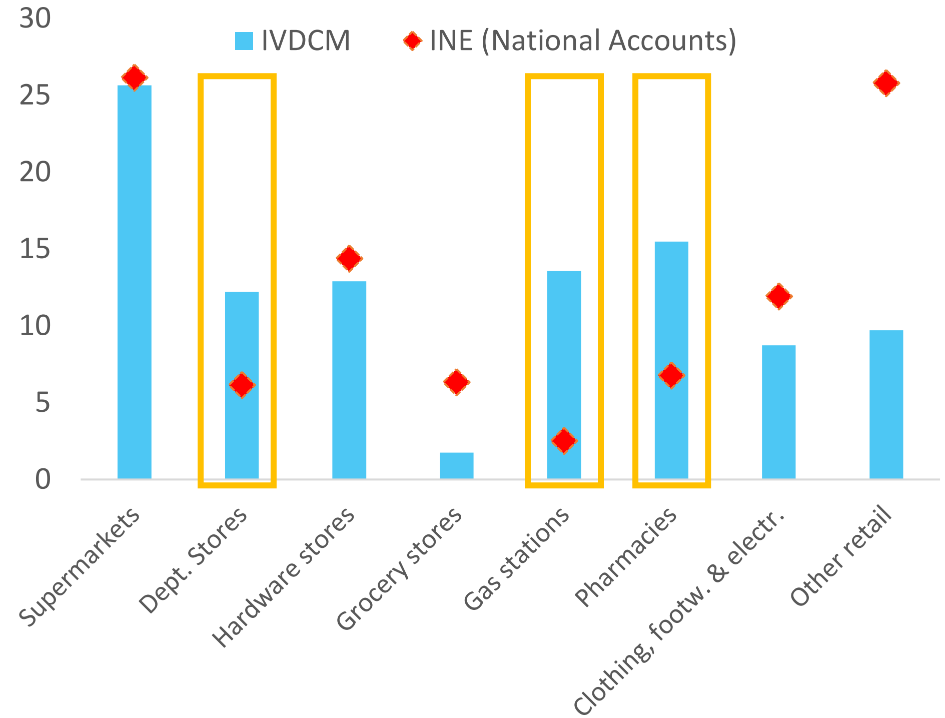
(share of total, percentage)



(*) Firm size calculated from annual sales according to SII definitions. Universe of firms from VAT declarations.

Weights by retail trade industry

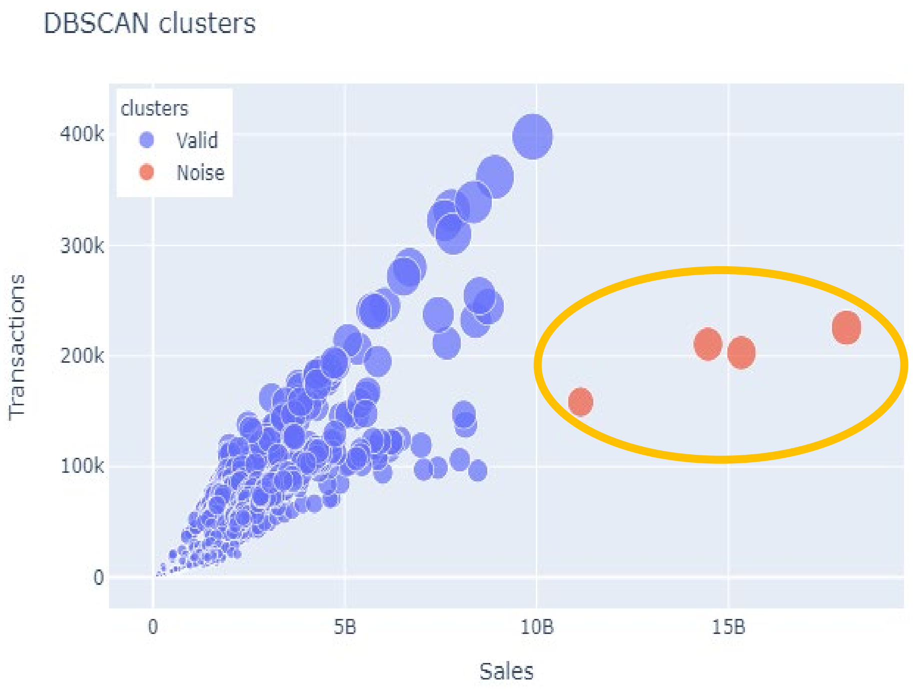
(share of total, percentage)



Challenges related to data quality:

Outlier detection

- We use the two dimensions available in the dataset, **number of transactions** and the **ammount of sales**, to form a distribution for **each firm**.
- Then, we apply **DBSCAN** which is an unsupervised clustering algorithm specially designed to **identify noise**.
- The algorithm identifies **dense regions** based on Euclidean distances. Any observation outside the valid clusters will be **marked as an outlier**.
- With new incoming data, the clusters are re-evaluated so that **new, valid behaviours** of firms can be considered normal.

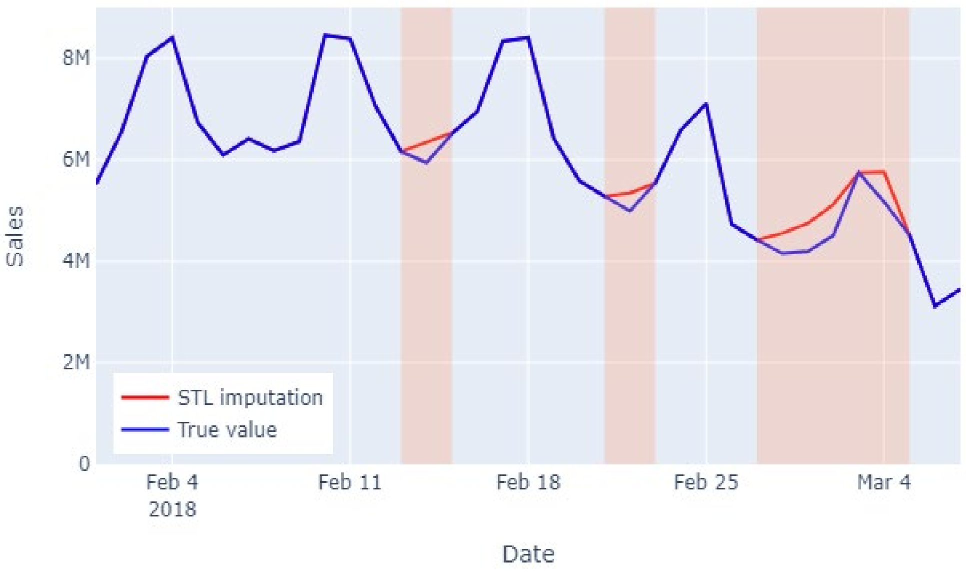


Challenges related to data quality:

Missing value imputation

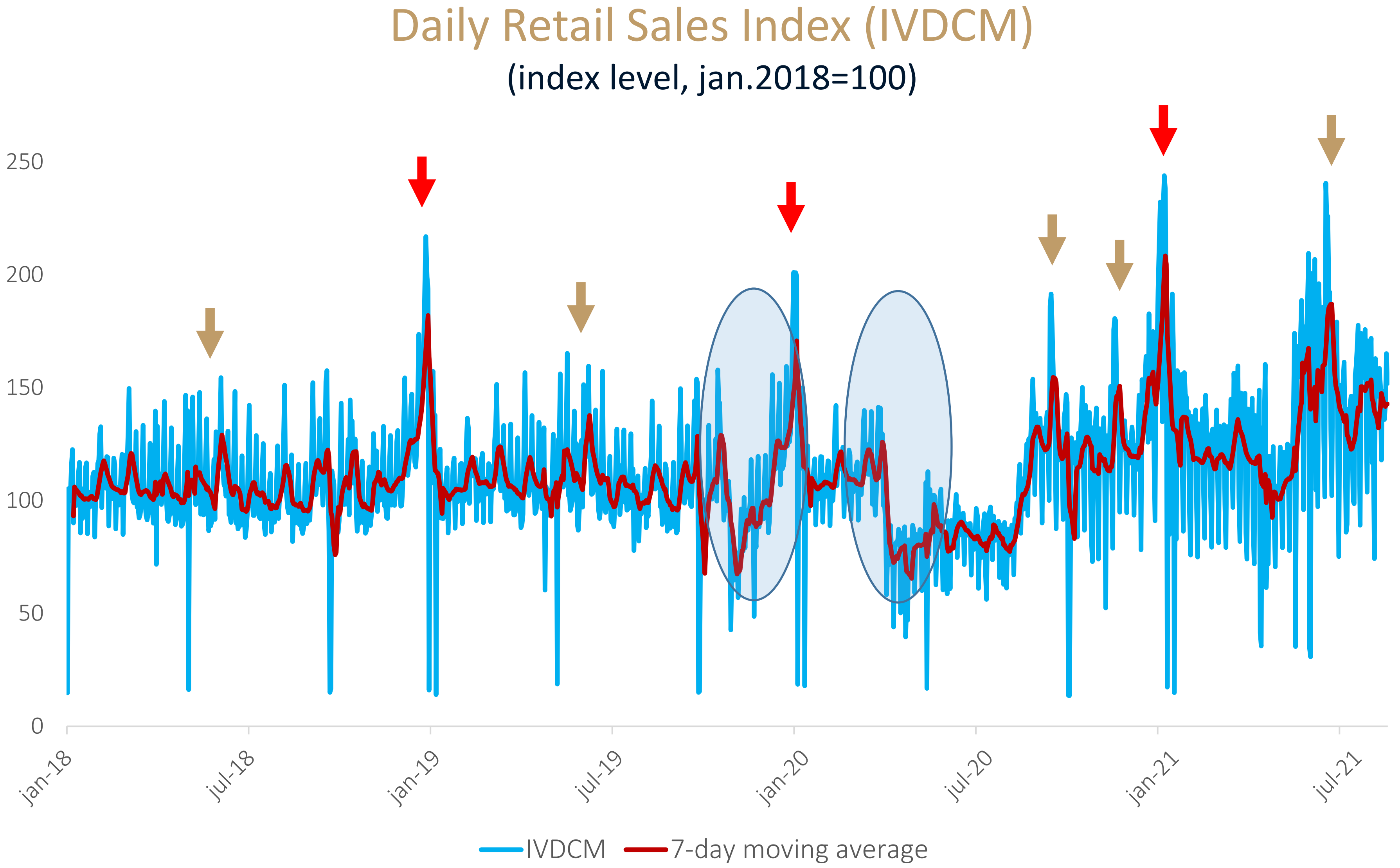
- Both true missing values and detected outliers are considered for imputation.
- To preserve the unique weekly seasonal pattern of each firm we use the **STL decomposition algorithm**.
- This is a **non-parametric** model from the econometrics toolbox that uses **local regressions** to obtain the seasonal and trend components, operating in the presence of missing values.
- STL **isolates the seasonal component**, so that the seasonally-adjusted series can be **interpolated** in the missing days. Then, the seasonal pattern **is added back** to obtain the final imputation.

Missing value imputation with STL decomposition



III. Results

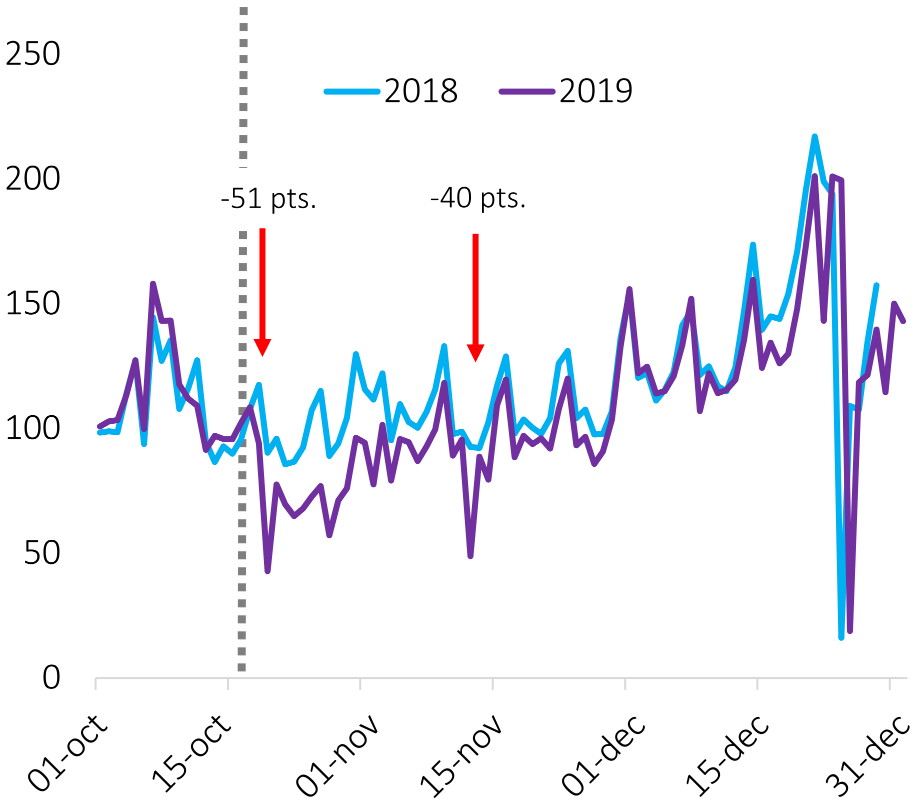
Daily Retail Sales Index (*IVDCM*) time series.



The daily index provides new insights about the magnitude of shocks on retail sales.

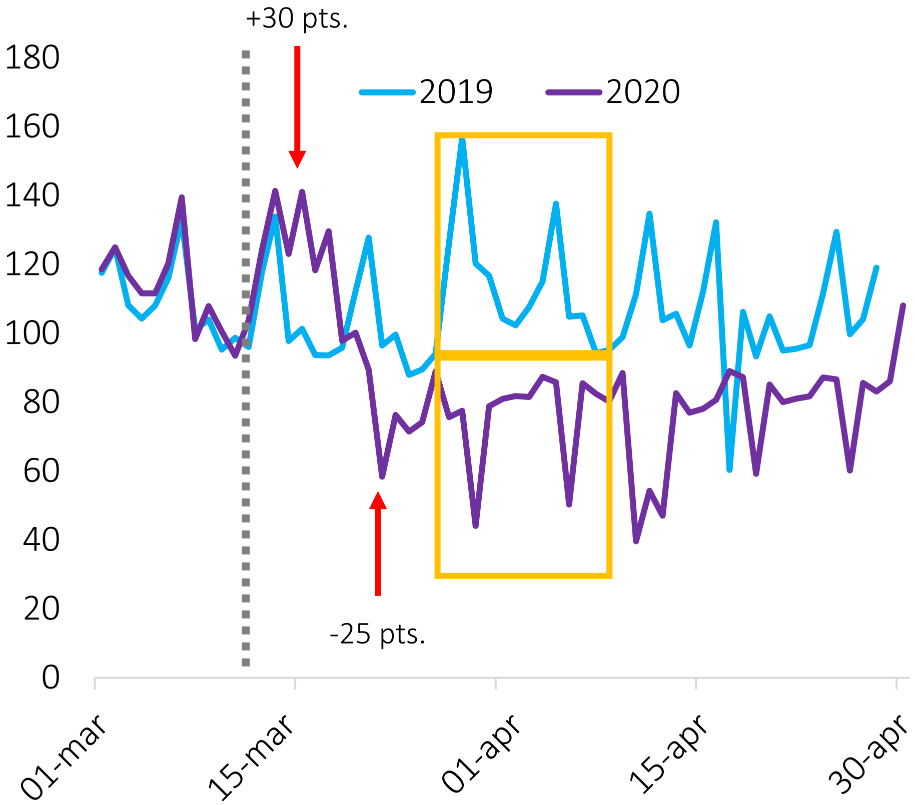
Social unrest (2019)

(index level, jan.2018=100)



Covid-19 crisis (2020)

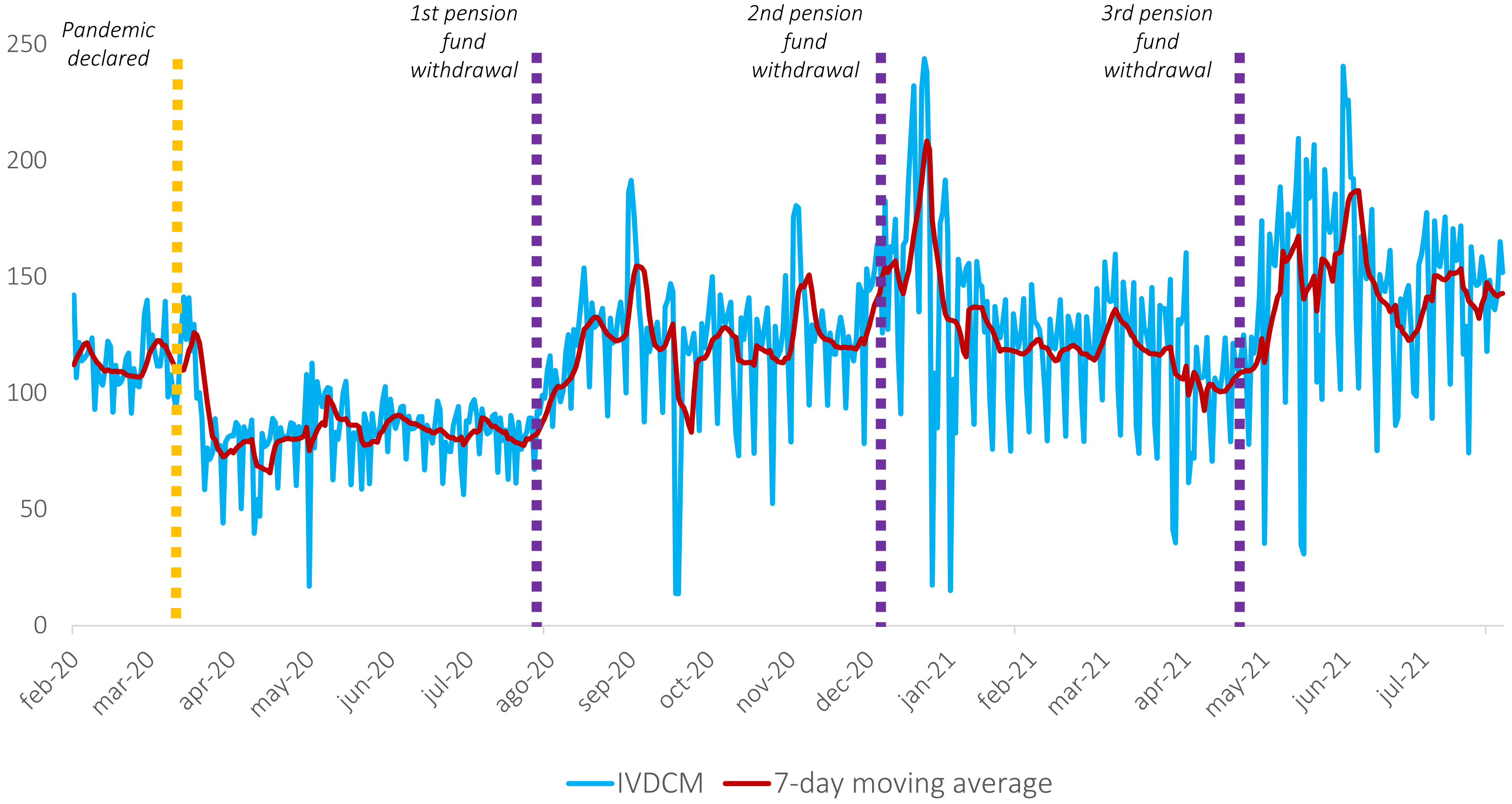
(index level, jan.2018=100)



The daily index provides new insights about the magnitude of shocks on retail sales.

Daily Retail Sales Index (IVDCM)

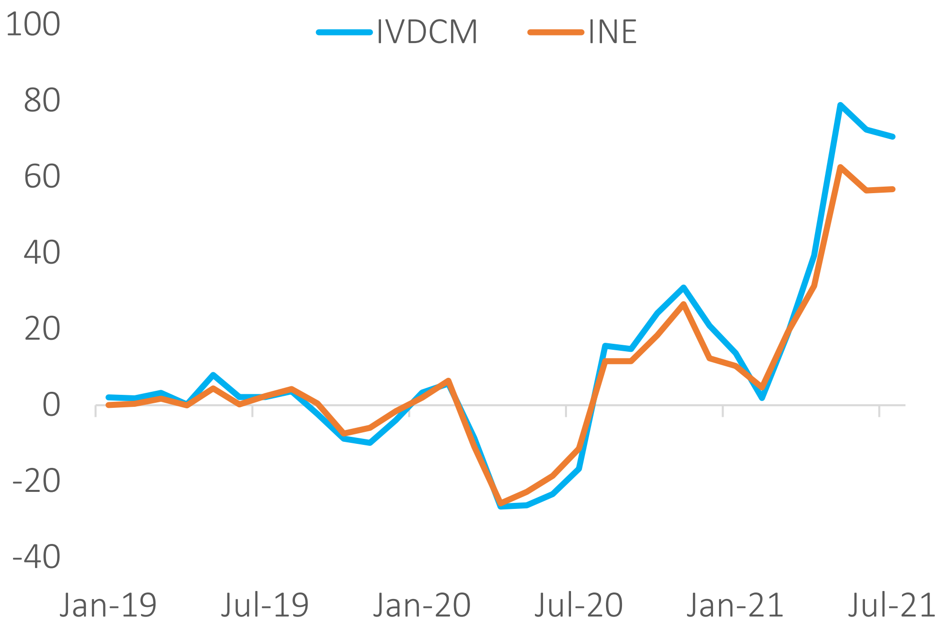
(index level, jan.2018=100)



IVDCM shows a positive and strong correlation with the official reference index (INE).

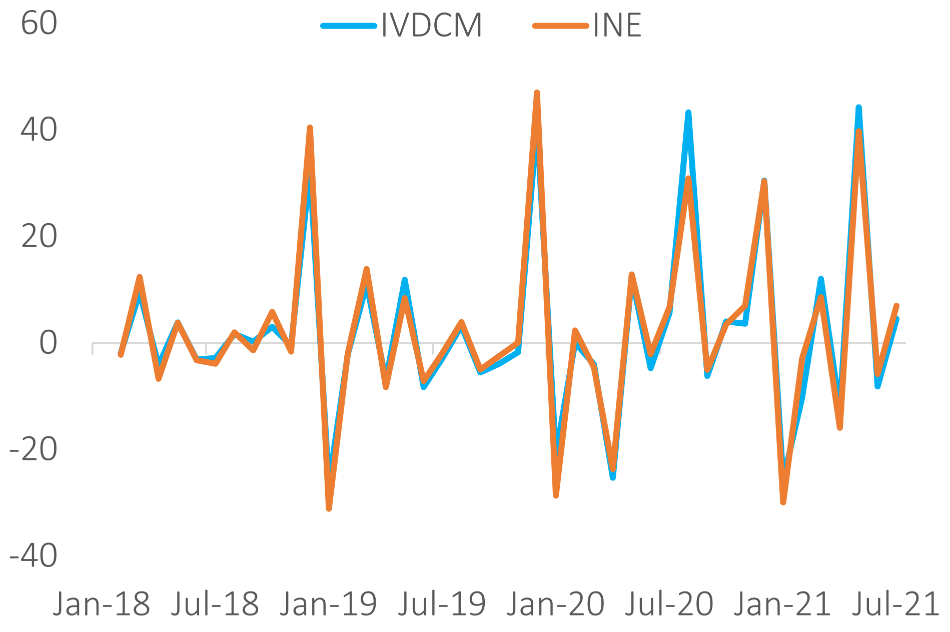
Retail trade

(YoY change, percentage)



Retail trade

(MoM change, percentage)

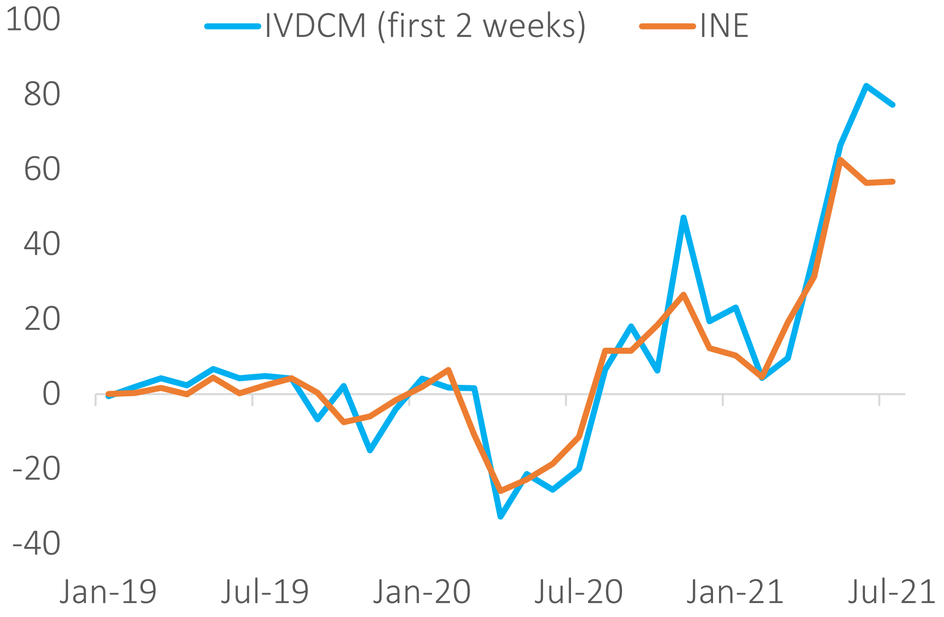


	YoY	MoM
Correlation	0.99	0.98
Median abs diff	2.7	2.3

Similarly, a first estimate of IVDCM with data for the first 2 weeks still shows high correlation with the reference index.

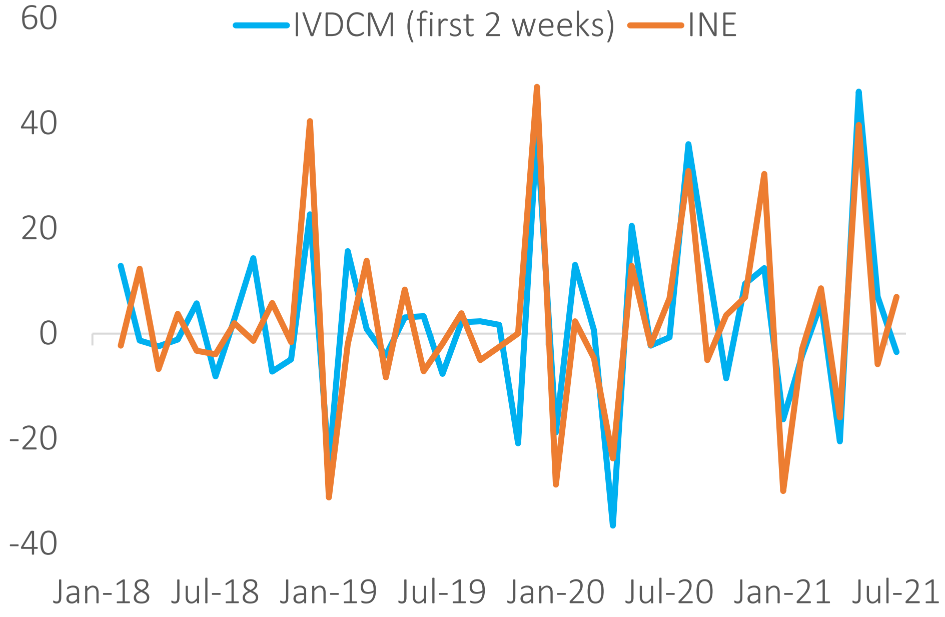
Retail trade

(YoY change, percentage)



Retail trade

(MoM change, percentage)



	YoY	MoM
Correlation	0.96	0.84
Median abs diff	6.0	7.5

At the Central Bank, results of this project have been used in Monetary Policy Reports for monitoring consumption.

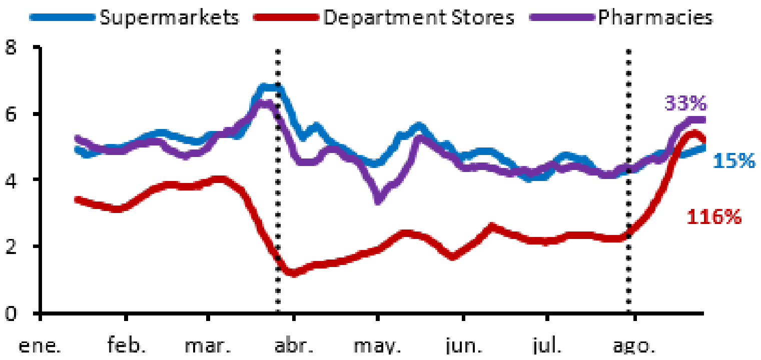
MPR - September 2020

Analysis of the impact of pension fund withdrawals

Figure II.13

Retail Sales with Electronic Receipts (*)

(two-week rolling average, billions of pesos)



(1) First vertical line marks the beginning of lockdowns in Santiago (March 26th). The second vertical line marks the publication of the law that allows pension fund withdrawals (July 30th). (2) Percentages show the change between July 30th and August 26th.

Source: Central Bank of Chile with SII data.

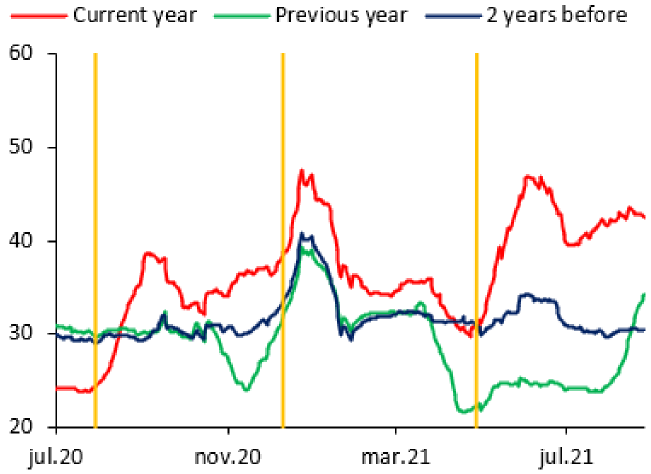
MPR - September 2021

Short-term dynamics of consumption

FIGURE III.4

RETAIL SALES WITH ELECTRONIC RECEIPTS (*)

(28-days rolling average, billions of pesos)



(*) Yellow vertical lines: pension fund withdrawals.

Source: BCCH with data from SII.

Future steps.

- This project is part of the **Experimental Statistics Initiative** of the Central Bank of Chile.
- Other projects (in progress) that are part of this framework: quarterly **business demographics**, distrubution of **household income**, consumption and wealth and **regional activity** indicators.
- With regards to IVDCM, we expect to **publish the index** in the coming months.
- We look forward for our user’s **feedback**!

High frequency tax information for tracking retail: challenges and applications

Joaquín Pérez

Statistics Division

IV Statistics Conference: “Post-pandemic Statistics”, Central Bank of Chile. September 28th 2021.