Invoices rather than surveys: using ML to build nominal and real indices

Pablo Acevedo
Central Bank of Chile

Emiliano Luttini
World Bank

Matías Pizarro
PUC-Chile

Dagoberto Quevedo
Central Bank of Chile

Marco Rojas
Central Bank of Chile

3rd IFC and Bank of Italy Workshop on Data Science in Central Banking
October 19, 2023. Rome, Italy

Disclaimer

The views are those of the authors and do not necessarily reflect the views of the Central Bank of Chile or its board members

Electronic Invoice

- The electronic invoice document is a mandatory digital document that reports the transactions between firms.
- The Central Bank of Chile obtains every day all documents issued in the last 24 hours ($\approx$ 3 millions).[1]
- Each document reports a free/unstructured text that describes each good or service in the transaction.
- However, it is *not standardized* to a product or service classification.

[1] Currently we have approximately 8,000 millions of documents with 19,000 millions free/unstructured text.

Example of an electronic invoice document

AGROSUPER COMERCIAL

AGROSUPER COMERCIALIZADORA DE ALIMENTOS LIMITADA

GIRO: Mayorista de Aves, Cerdos, Vacunos, Cocinas, Balnearios Congelados, Verduras, Hortalizas y Frutas Congeladas, Comercialización de Vinos y Licores.
CASA MATRIZ: La Estrella 401 of. 7 Sector Punta de Cortés - Rancagua - Fono (72) 233940
DIR. CABLEG.: FAX: 5306165 - SANTIAGO CHILE - CASILLA 277 - STGO. / 23 MAIPU
SERVIFONO: 600 600 6001

AGROSUPER

R.U.T.: 79.984.240 - 8

FACTURA ELECTRONICA

Nº 74

S.I.I. - RANCAGUA

SEÑOR(ES):	Fernanda Chacana Serrano	R.U.T.:	14.205.748-2
DIRECCION:	Punta Cortes s/n	150966 595985 0 1	San Pablo
GIRO:	Almacén		
COMUNA:	Pudahuel	CIUDAD:	Prov Santiago
VENDEDOR:	16030-Hector Iver Moraga Carreño	DIR. ORIGEN:	
FECHA:	05/05/2003		San Pablo N°9500,Pudahuel-Prov Santiago
CONDICION DE VENTA:	Efectivo	FECHA DE VENCIMIENTO:	05/05/2003
LOCAL:	Punta Cortes s/n		Pudahuel Prov Santiago

PLU	CODIGO	PRODUCTO	ENV.	EST.	UNID.	KILOS	PRECIO	VALOR
	04-01-02-062	(Super)Mortadela Jamonada-Fresco	09-09	1	1	3,00		
	04-01-02-064	(Super)Mortadela Salch. Cerveza C-75-Fresco	09-09	1	2	3,30		
	04-01-06-186	(Super)Arrollado lomo con ají-Fresco	09-09	1	1	2,60		
	04-04-02-068	(Super Pollo)Mortadela jamonada de pollo-Fresco	09-09	1	1	3,03		

Figure 1: Example of electronic invoice document. Source: Servicios de Impuestos Internos (SII).

Motivation

- Challenge:

  → We do not know **what** products are actually being sold.

- What we do:

  1. Use cutting-edge machine learning techniques. to classify transactions into product categories.
  2. Measure the economy in real time (e.g., consumption, investment).
  3. Study the differential effect the pandemic (and lockdowns) had on different consumption goods.

Roadmap

1. Classification approach
   - Strategy
   - Classification model
   - Quality metrics

2. Applications

Classification approach

Strategy

To standardize the free-text description, we do the following:

1. Consider two classifiers (COICOP, CUP):
   - Classification of Individual Consumption According to Purpose (COICOP): 303 classes.
   - Unique Code of Chilean Products (CUP): 290 classes.
2. Build a training sample using a human-entity to label items to this classifiers. Detail
3. Train a supervised multi-class text classification model to extrapolate learning to the rest of documents.

Training sample composition[2]

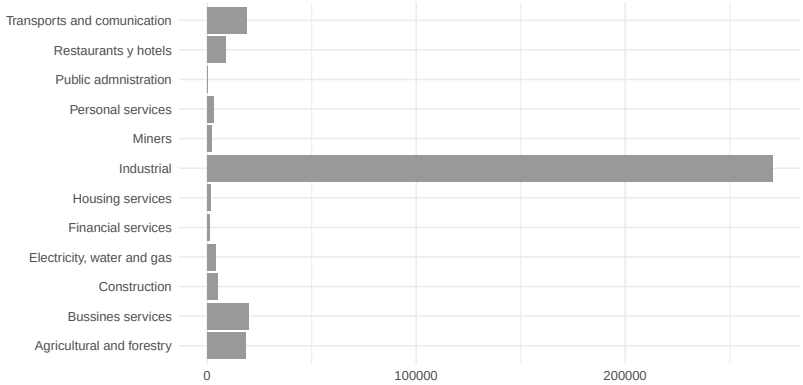


Figure 2: Training sample composition by Division on Unique Code of Chilean Products.

[2]Until December 2022

Expected output

Free/unstructured text	Class code	Class name
Fairlife 2% Reduced Fat Ultra-Filtered Milk, Lactose Free, 52 fl oz	011412	Liquid milk
Fairlife 2% Chocolate Ultra-Filtered Milk, Lactose Free, 52 fl oz	011412	Liquid milk
Great Value 2% Reduced Fat Milk, 128 Fl Oz	011412	Liquid milk
Great Value Fat Free Milk, Gallon, 128 fl oz	011412	Liquid milk
Crystal Whole Vitamin D Milk, Gallon, 128 fl oz	011412	Liquid milk
Horizon Organic Whole Shelf-Stable Milk Boxes, 8 Oz., 12 Count	011412	Liquid milk
Minute Instant Brown Rice, Light and Fluffy, 28 oz	011110	Rice
Minute Instant White Rice, Light and Fluffy, 14 oz	011110	Rice
Ben's Original Ready Rice Fried Flavored Rice, Easy Dinner Side, 8.5 OZ Pouch	011110	Rice
Success Rice Boil-in-Bag White Rice, Family Size, 32 Oz	011110	Rice
Clorox Splash-Less Liquid Bleach, Regular (Concentrated Formula) 77 Ounce	056114	Cleaner
Clorox Clean-Up All Purpose Cleaner with Bleach, Spray Bottle, 32 oz	056114	Cleaner

Word embedding

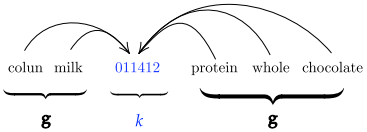
- Vector representation of words
  - Encoding: words that are close in the vector space are also similar in context to labelcode for class k.
- Many pre-compilations are available for Spanish and could be used in other contexts
  - However, it does noy apply in our case:
    $\Rightarrow$ We build a word embedding model is using words from training sample.
- In order to reinforce the fit of model, other features of the transaction are added as complements of the unstructured text:

$$\mathbf{g} = \{g \cup \langle b \rangle \cup \langle s \rangle\}, \tag{1}$$

g = standardized text after to applied a text cleaning strategy, Detail
b = brand,
s = sector that issued the text.

Definition of classification model

- We use FastText (Joulin *et al.*, 2016) as the classification model (similar to the CBoW (Mikolov *et al.*, 2013)), where the middle word is replace by a labelcode for class k:



- The softmax function f gets the probability of observing label for class k given a paragraph $\mathbf{g}$:

$$f(k, \mathbf{g}) \equiv \mathbb{P}(k \mid \mathbf{g}) = \frac{e^{h_{\mathbf{g}}^{\top} \mathbf{v}_k}}{\sum_{l=1}^{K} e^{h_{\mathbf{g}}^{\top} \mathbf{v}_l}},$$

where, $\mathbf{u}_w \in \mathbb{R}^D$ = vector representation of word w in the D documents in training sample,
$\mathbf{v}_k \in \mathbb{R}^D$ = vector representation of class k,
$h_{\mathbf{g}} \in \mathbb{R}^D = \sum_{w \in \mathbf{g}} \mathbf{u}_w$, paragraph features of $\mathbf{g}$.

Fine tuning and execution conditions

- The fastText library was downloaded from GitHub[3] repository and compiled.
- For each classifier we run an exhaustive hyperparameter grid search using the *automatic hyperparameter optimization*[4] option (`autotune`), running for 12 hours, getting the set of hyperparameters for build the final model.
- All execution were run on a machine with the Windows Server 2022 operating system, Intel Xeon Gold 6126 2x2.60 Ghz processor and 1.5 TB RAM, under Python 3.11. `scikit-learn` package was used for run the quality metrics.

[3] github.com/facebookresearch/fastText
[4] fasttext.cc/docs/en/autotune.html

Quality metrics

- A repeated stratified q-fold cross-validation is done, using $q = 5$ and stratifying by class k, then repeated 10 times.
- For each repeated cluster $q \times k$, we compute the metrics of Precision, Recall and F1.
  Detail
- Results:

Table 1: Quality metrics by classifier.

Classifier	Precision	Recall	F1
COICOP	0.955	0.954	0.954
UCP	0.936	0.935	0.935

Notes: A weighted average by number of text in class k is shown for each quality metrics by classifier.

Applications

1. Durable and non-durable consumption

- Our data is firm-to-firm: not consumption of households
  - We solve this by using *purchases* of the retail sector.
- We map the product categories into durable and non-durable consumption
  - Then, we sum across all transactions for durable/non-durable to obtain nominal aggregates.
  - Obtain real aggregates using official deflators.

$\Rightarrow$ Build real-time consumption of durable and non-durable

  - Comparable to National Account series.

- We can do this for finer groups of products.

Durables and non-durables series

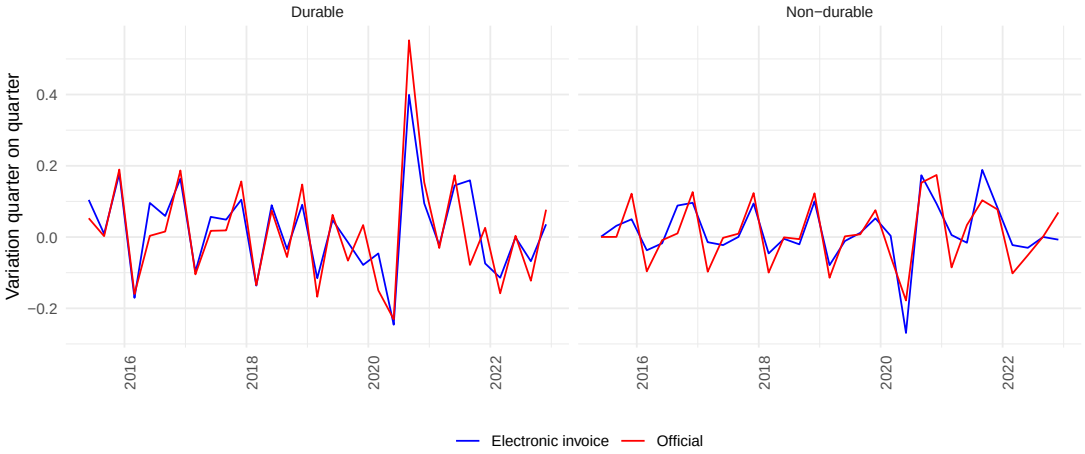


Figure 3: Volumen variation: quarter on quarter delta log changes. Authors own calculation based on Central Bank of Chile data.

2. Consumption pattern before and during lockdown

What we do

- We assess the consumption pattern before and during lockdowns during the COVID-19 pandemic.

How we do it

- We build series from the retail sector purchases.
- Goods are identified using the COICOP classification.
- An empirical approach is proposed to consider purchases variations with respect to the first week of the lockdown. Detail

Heterogenous effects of lockdown on consumption

By macrodivision product

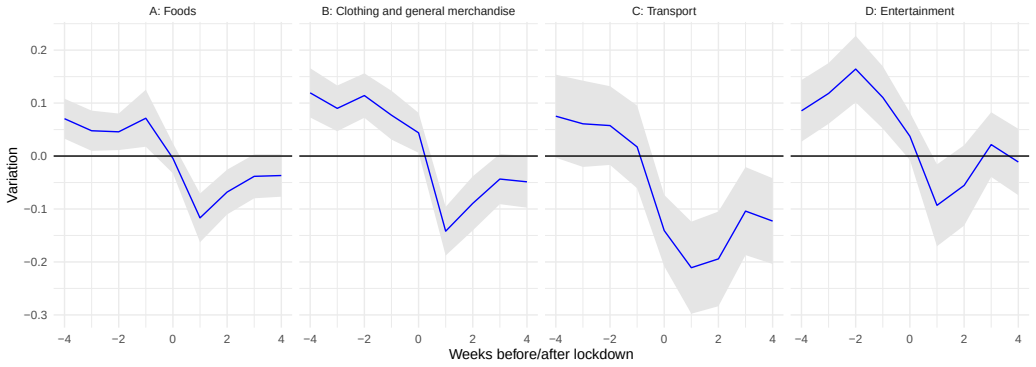


Figure 4: Purchases variations per week before or during lockdown for retail sector by division product. This figure report $\sum_{s \in S} \alpha_s$ of estimate regression define in Equation 4 of the empirical approach. The gray area denote a confidential interval 5-95%.

Conclusion

1. We propose a methodology to go from raw description of goods sold to standardized product categories.
2. We use frontier techniques in machine learning together with other observables from tax documents to carry out the classification.
3. We apply this to build real-time series of consumption (and investment), and to study the effect of lockdowns on different types of consumption.

Appendix

Text cleaning strategy

Return

The text component or features are standardize, removing elements into the text whose inclusion not contribute to the classification process. The following operations are applied:

1. Transform text to lowercase.
2. Standardization to Unicode format NFKD for transform to ASCII encoding, e.g. replace accents.
3. Remove any character distinct to $[a - z]$.
4. Remove stopwords.
5. Apply stemming.
6. Remove words with a length less than two.
7. Replace white space with a length greater than one, and any white spaces in start o end of the text.

Training sample

Return

- A selection criteria statistic is defined for each text $i \in G_s$ in sector s:

$$\gamma_{si} = \frac{1}{3}\left(\frac{m_i}{\sum_{j \in G_s} m_j}\right) + \frac{1}{3}\left(\frac{e_i}{\sum_{j \in G_s} e_j}\right) + \frac{1}{3}\rho_i^s, \tag{2}$$

where,

m_i = total revenue,
e_i = number of events or reports,
ρ_i^s = monthly normal probability occurrence over time,
G_s = set of unique free texts in sector s.

- n free text in sector s with the highest value of γ_{si} are selected.
- A person interprets the text and labels items to a class that represents the good or service.

Quality metrics

Return

- For each run cluster q and class k, we calculate the number of true positives (T_p), of false positives (F_p), and of false negatives (F_n).
- Precision (P): measure of result relevancy,

$$P^{qk} = \frac{T_p^{qk}}{T_p^{qk} + F_p^{qk}},$$

- Recall (R): measure of how many truly relevant results are returned,

$$R^{qk} = \frac{T_p^{qk}}{T_p^{qk} + F_n^{qk}},$$

- F1-score: harmonic mean of precision and recall,

$$F1^{qk} = 2\frac{P^{qk} \times R^{qk}}{P^{qk} + R^{qk}}.$$

Empirical Approach

Consumption pattern before and during lockdown

Return

Weekly purchases change (in relation to the ones done in the first week of lockdown) is determined as:

$$\Delta \log X_{t,c} = \log X_{t,c} - \log X_{\rho(t),c}, \tag{3}$$

where $X_{t,c}$ is the total purchases in the municipality c and period t and $\rho(t)$ is the week of the closest lockdown start. This is incorporated on the following model:

$$\Delta \log X_{t,c} = \lambda_c + \sum_{s \in S} \alpha_s \mathbb{1}\{s(t) = s\} + \sum_{w \in W} \beta_w \mathbb{1}\{w(t) = w\} + \epsilon_{t,c}, \tag{4}$$

where,

S = set of weeks before or during lockdown, $\{-4, \ldots, 0, \ldots, 4\}$,
W = set of weeks number in month, $\{1, \ldots, 5\}$,
$s(t)$ = weeks before or during lockdown in t, $s(t) \in S$,
$w(t)$ = week number in month of t, $w(t) \in W$.