Irving Fisher Committee on
Central Bank Statistics

BIS

IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks"

Property classification with administrative data: The case of the Chilean household price index[1]

Juan José Balsa and Javiera Vasquez,
Central Bank of Chile

[1] This contribution was prepared for the IFC Satellite Seminar held at the ISI 64th World Statistics Congress, co-organised with the Bank of Canada in Ottawa, Canada, on 15 July 2023. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Canada, the BIS, the IFC or the other central banks and institutions represented at the event.

Working paper

Property classification with administrative data: The case of the Chilean House Price Index[1]

Juan Jose Balsa[2]
Javiera Vásquez[3]

Department of Microdata
Division of Statistic
April 2023

Summary

Working with administrative data can be particularly challenging when it is incomplete or lacks critical information. In this paper, we present a specific case study involving the Chilean House Price Index, that is computed with a mixed stratification method where data classification into apartments and houses is necessary for the robustness of the method. In compiling this index, the Central Bank of Chile uses administrative information from the SII[4] on all effective transactions of residential properties.

We aim to provide a step-by-step approach to data classification, including testing various algorithms and optimizing the best ones for the data. Our approach helps us to classify over 3 million observations in a comprehensive and fast way. The paper discusses the House Price Index methodologies, the different method for the classification of the property, the classification algorithms used, and provides codes and insides of the challenges we encountered during the process.

[1] We thank Maria Fernanda Castillo for their contributions to this work.
[2] Juan José Balsa (jbalsa@bcentral.cl), analyst of the financial microdata group of the BCCh, Msc. in Economics from the University of Groningen and Msc. in economics analysis from the University of Chile.
[3] Javiera Vasquez (jvasquez@bcentral.cl), Head of the financial microdata group of the BCCh, Msc in Economics from the University of Chile.
[4] From the spanish "Servicio de Impuestos Internos" (Tax Administration Office)

Contents

Introduction

Housing represents a significant source of wealth and debt for households globally. For the case of Chile in 2021 the 63.7% of household assets correspond to real estate assets, and 54.3% of debt is constituted by mortgage debts[5]. Thus, variations in housing prices can have significant implications for household finances, with repercussions in the financial system, consumption, and the national economy.

Therefore, it is reasonable to continuously monitor the evolution of housing prices due to their effects on the country's financial stability. In fact, after the severe consequences of the subprime mortgage crisis, there has been increased interest in the systematic analysis of price and activity indicators in the real estate sector.

In addition, the increases in monetary policy rates during 2022, caused by the global high inflation scenario that affects debt and access to credit for households and companies in this sector, make it highly relevant to review and continuously improve this indicator.

Different methodologies for constructing housing price indices have been documented in the literature, and the choice of the most appropriate one will depend mostly on the data available and the use of the index . For example, the Eurostat handbook indicates that if the purpose is to estimate the dynamics of prices to track housing market inflation, the optimal way is to gather information on current transaction prices. On the other hand, if the objective is to estimate the balance of a country's wealth, information on housing sales should be supplemented with information on the stock of properties that are not for sale. However, in practice, the availability of information determines the methodology to be used.

Within the mission of the Central Bank of Chile (BCCh) is to ensure financial stability, and analyzing the real estate sector is crucial, where a price index is a central tool. Therefore, starting in 2014, the BCCh developed the Housing Price Index (IPV), which is currently calculated on a quarterly basis and published in the Statistical Database. The main objective of this index is to study the behavior of prices for urban residential properties in the Chilean market.

The IPV[6] is calculated using the mixed stratification method[7], where the 27 strata are defined based on seven geographic zones, housing condition (new-used), and housing type (house or apartment)[8]. The general IPV, which represents transactions throughout the national territory, is published along with specific indicators by geographic zone, property condition, and housing type. In compiling this index, the BCCh uses administrative information from the SII[9] on all effective transactions of residential properties.

A central part of the process to ensure the correct generation of the IPV is the stage of cleansing and imputing administrative records sent by the SII in both the F2890 form and

[5] Chilean Household Financial Survey available on Encuesta Financiera de Hogares (EFH) - Banco Central de Chile (bcentral.cl)
[6] From the spanish "Índice de Precios de Vivienda" (Household Price Index)
[7] For more detailed information, please refer to Section two of this document.
[8] The strata of new houses in the city center of Santiago has very few transactions so it is eliminated.
[9] From the spanish "Servicio de Impuestos Internos" (Tax Administration office)

the Real Estate Registry (CBR). These cleanings are complemented with the creation of new variables that are not in the records but are requiered for the generation of the strata, such as classification by housing type, and the condition of the property (new or used). Given the complexity of the previously described process and in line with delivering more and better statistical information, the BCCh during 2021 conducted a review of the cleansing and imputation processes of the information provided by the SII, along with the incorporation of new sources of information to enrich the IPV calculation. The review led to a series of improvements in the calculation of the indicator, reducing its volatility. The objective of this document is to present one of the key stages of the process of creation of the IPV, the property type classification among houses and apartments.

Classifying administrative data can be challenging, and in the case of the IPV, the challenge was to differentiate between houses and apartments. Various techniques were used for this classification, including manual classification, regular expressions, close neighbors, text mining, and machine learning algorithms such as logistic regression, random forest, k-means, k-nearest neighbors, and support vector machines.

The first section of the document describes house price indexes and their characteristics, main methodologies, and some international experiences. The second section presents a summary of the IPV methodology. The third section discusses classification algorithms and their evolution in the literature. Finally, in the fourth section, we make a zoom in the house type classification, the different approaches, we present a ready to use toolkit for others to use and some results.

House Price Indexes

A price index is an indicator that reflects the evolution over time of the price level of a segment of the economy. According to the Eurostat manual (2013), for a price index to be useful in the decision-making of various economic agents, it must meet the following characteristics: i) transparent methodology based on objective and documented criteria, ii) observed prices from actual transactions, and iii) easy access for user monitoring.

Housing is a capital good that has certain characteristics that make it difficult to construct a price index for this market. As the Eurostat Housing Price Index Manual (2013) states, houses are fundamentally heterogeneous since there are no two identical houses, either due to their physical characteristics or their location. This implies that the price of one property is not necessarily a good predictor of the price of another.

In addition to this, there is the problem of temporality, as it is not possible to exactly equate two properties due to the depreciation they undergo and the repairs or changes that owners may make over their useful life.

The house price index (HPI) consists of an indicator that measures the change in the price of residential real estate over time. It is often used as an indicator of the health of the housing market and can provide information on the state of the economy.

According to Himmelberg, Mayer, and Sinai (2005), the HPI is typically calculated from price data on actual home transaction, considering factors such as location, size, property

type, and overall condition. These data are then compiled into a single figure, representing the average change in housing prices in each area. This provides a useful tool for tracking trends and changes in prices over time.

The relevance of the HPI in the real estate sector can be diverse. For example, for homeowners, the HPI is used as an indication of the current value of their property, which can be useful to make decisions about selling or borrowing. Englund, Hwang, and Quigley (2002) note that the HPI can help homeowners assess the equity they have in their homes and make informed decisions about their financial future. For real estate agents, on the other hand, the HPI can help them set sale prices and predict housing demands. Meanwhile, for investors the HPI is a key indicator for assessing the performance of real estate assets and identifying potential growth or diversification opportunities. Case, Quigley, and Shiller (2013) note that the HPI provides a useful benchmark against which investors can compare their portfolio performance and make informed investment decisions.

In addition, the HPI is also important for policymakers, as it can help them identify and respond to changes in the housing market. For example, if the HPI shows that house prices are rising at a fast rate, policymakers may consider implementing measures to control the credit market, like raising interest rates. On the other hand, if the HPI indicates that house prices are falling, policymakers may consider introducing stimulus measures to support the housing sector and thus the economy.

I. Main indexes methodology

I.1 Mean or Median

The simplest analysis of housing price trends is based on measuring the central tendency of the distribution of prices of properties transacted in a particular period, using the mean or median. A significant disadvantage of simple indices (mean or median) is that they produce noisy estimates of price variations[10], since the set of homes sold in a period is often small and not necessarily representative of the stock of homes. Thus, changes in the composition of properties sold affect the sample of properties on which the mean or median is being calculated. Additionally, simple indexes are subject to biases when there are changes in the quality of properties over time or when certain types of homes are sold more than others.[11]

I.2 Stratification

The stratification method, also known as "composition adjustment," is based on the creation of strata or groups of transactions that aim to neutralize variations in the composition of properties sold. [12]

Stratification involves breaking down the total sample of transactions into a series of sub-samples or strata. Then, for each stratum, the price index based on the mean or median is obtained to calculate the aggregated housing price index as a weighted average of the indexes of each stratum. That is:

---

[10] Eurostat manual (2013). Pg 12.

[11] Prasad, N. L., y Richards A. (2008). Improving Median Housing Price Indexes Through Stratification.

[12] Moreover, stratification allows for the calculation of sub-indices for different segments of the housing market.

$$P^{0t} = \sum_{m=1}^{M} \omega_m^0 P_m^{0t}$$

Where P_m^{0t} is the index of the stratum m that compares the average price of the period t is the period with base 0, and ω_m^0 denotes the weight of the stratum m.

The formation of the strata is based on certain variables, such as geographic location, type of housing, housing condition, property size, etc. In this way, changes in the composition of housing between different periods are controlled, but not within each stratum. The definition of the stratification variables is key and should allow for the construction of the most homogeneous strata possible. It is important to note that if variables are not being considered in the stratification, changes in the composition according to those variables will not be controlled for.

Regarding the value of the weights, it depends on the type of index being estimated. If the main objective is to monitor the price variation of the stock of housing, then weights based on the existing stock of housing should be used, i.e., the value of the stock corresponding to each stratum. On the other hand, if the goal is to measure the variations in prices of the housing transacted in the market, weights based on sales should be applied.

The main advantages of the stratification method are that it is simple to implement and reproducible by any user who knows the stratification variables. It also allows for the construction of different sub-indexes within the housing market. However, this method assumes that the stratification variables are sufficient to control for all changes in composition.

If the stratification is too granular, the indexes calculated in each stratum can be very unstable due to the small sample size available in each cell. On the other hand, if the stratification is not granular enough, the composition variables will affect the indexes. Therefore, it is essential to achieve a balance in the number of variables used for stratification.

I.3 Repeat Sales

The name of this method indicates that it uses information about properties sold more than once. This way, it controls for the characteristics of the property, assuming that they do not change between periods and that price changes do not reflect changes in the quality or composition of the sample. However, implicitly, new properties for sale are excluded from an index constructed with this methodology.

To construct the house price index based on repeated sales, a simple regression model must be estimated that requires very few variables: the price and the sale date of the property. Thus, if there is a sample of properties sold in a period of time t=0,...,T, with their respective transaction prices, the linear regression model to be estimated is:

$$ln\left(\frac{p_n^t}{p_n^s}\right) = \sum_{t=0}^{T} \gamma^t D_n^t + \mu_n^t$$

Where p_n^t is the sale price of the property n in the period t; D_n^t is a dummy variable that take value 1 in the period of re-selling n, -1 in the period of the previous sale and 0 in any other case and μ_n^t is the error term.

Thus, the repeat sales index that goes from period 0 to period t is obtained as follows:

$$P_{VR}^{0t} = \exp(\hat{\gamma}^t)$$

While this method is simple and easy to implement, a problem with the repeat sales method is that the property sold at two different points in time is not necessarily identical due to factors such as depreciation and renovations. Eurostat (2013) recommends excluding properties that are sold quickly or that are not resold for prolonged periods from the calculations, that is, limiting the observation period for resales.

I.4 Hedonic Regression

The hedonic regression method assumes that the heterogeneity of goods can be represented through a series of observable attributes or characteristics. In the case of real estate properties, these attributes are related to the structure and location of the property. In this context, demand and supply for properties implicitly determine the marginal contributions of characteristics to property prices, and regression techniques can be used to estimate these marginal contributions.

The key assumption for the use of the hedonic regression method is that the price of property n at time t, P_n^t, is a function of a number K of characteristics that are measured through the variables z_{nk}^t:

$$P_n^t = f(z_{n1}^t, z_{n2}^t, \dots, z_{nK}^t, \varepsilon_n^t) \quad con \quad t = 1, \dots, T.$$

Where ε_n^t is a stochastic error term

To estimate the marginal contribution of each characteristic using the regression method, it is necessary to parameterize the function that relates the price of the property to its characteristics. The two most used hedonic specifications are:

- linear model:

$$P_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t$$

- Log-linear model:

$$\ln P_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t$$

In both specifications, it is allowed for the coefficients associated with the characteristics to vary over time, which is consistent with the idea that market conditions change in time;

however, these variations are gradual, allowing for the assumption of constant coefficients over time. Thus, the linear logarithmic model can be reduced to:

$$\ln P_n^t = \beta_0^t + \sum_{k=1}^{K} \overline{\beta_k} z_{nk}^t + \varepsilon_n^t$$

Therefore, if we have samples of transactions for each period, we could estimate the above model for each period separately, or combine all the periods and include dummy variables for those periods:

$$\ln P_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \overline{\beta_k} z_{nk}^t + \varepsilon_n^t$$

An important aspect of hedonic models is that, like any regression model, all relevant characteristics of the property that determine the price should be included as explanatory variables. Otherwise, a bias in the housing price index is introduced due to omitted relevant variables.

Then, the coefficients of time-related dummy variables are used to obtain an indicator of the variation in housing prices, controlling for the characteristics[13]:

$$P^{0t} = \exp(\hat{\delta}^t)$$

II. Prices Indexes around the world

At an international level, both public and private organizations have constructed housing price indexes with the objective of providing useful information for research and decision-making. These indexes use different sources of information and methodologies, depending on the quality and availability of data. Below are some examples of Housing Price Indexes grouped according to the three methodologies mentioned in the previous section.[14]

Hedonic Regression

In Sweden, the SCB (Sweden Central Bank) uses an estimated value of properties based on hedonic regressions. They use a method that combines the use of transaction prices and appraisal prices. The properties are divided into 12 strata (according to price) based on estimated values (appraisals). The weights are calculated based on the stock of housing in each stratum, which allows for the calculation of the average sales prices in different strata and then the estimation of an index.

In addition, another index exists in Sweden: The Nasdaq OMX Valuegard-KTH, which, unlike the former, is calculated by a private company and is commercialized. It is based

[13] For further insight into other methods for obtaining the housing price index using hedonic models, see the Handbook on Residential Property Prices Indices (RPPIs), Eurostat (2013).
[14] Further information on Owusu-Ansah, A. (2018)

solely on actual transactions of houses and condominiums. This index is calculated for three metropolitan regions and uses a hedonic regression method with time variables.

In Germany, the Hypo Real Estate Index (HyperIndex) has been computed since 2008. Its main purpose is to provide useful information on the real estate market, especially because the German Banking Act stipulates that financed housing transactions must be constantly monitored. This index uses data from the German banking association on real transactions. There is an index for houses and apartments, a national index is published, and there is a specific index for some cities.

Stratification

The Australian Bureau of Statistics (ABS) publishes a bi-annual index of housing prices for the 8 capital cities in Australia, as well as a weighted national average. The index is constructed using the mean and is compiled using a mixed stratification based on two characteristics: the long-term price level for the suburb where the property is located, and the socioeconomic characteristics of the neighborhood.

Repeat Sales

In the United States, the most recognized index is the S&P/Case-Shiller, based on the repeat sales method. The index monitors the growth rate of residential properties at both the national level (published quarterly) and for 20 metropolitan regions (published monthly). The data for the construction of the index are managed by Fiserv Inc., a leading provider of information technologies that receives information on home sale prices from multiple sources and then cross-validates them before including them in the database.

In Canada, Teranet (an international leader in electronic property registration) along with the National Bank of Canada, produces a publicly available price index. This corresponds to an independent representation of the growth rate of individual house prices. The index covers eleven metropolitan areas of Canada. Table N°1 summarizes some international experiences in the development of housing price indices:

Table N°1. International Experience in House Pricing Index (HPI)

Country	Institution	Frequency	Data	Source	Methodology	Coverage
Mixed Adjustment						
Australia	Australian Bureau of Statistics	Quarterly	Sample	State and Territory Land Titles Office/Real Estate Agents	Stratified Median	National Index + 8 capital cities
Greece	Bank of Greece	Quarterly	All transactions	Banks and credit institutions/ Real Estate Agents	Geometric Mean Stratified	National Index + Categories[15]
Turkey	Central Bank of the Republic of Turkey (CBRT) Turkey Statistical Institute (TurkStat)	Quarterly	Mortgage transactions	Commercial Banks and Real Estate appraisal companies	Stratified Median	National Index + Capital (all dwellings)
Hedonic						
Sweden	Sweden Central Bank	Quarterly	Real transactions of purchases of new dwellings and Purchases of existing dwellings	Real Estate Tax Assessment Register	Mixed Method: Hedonic Regressions + SPAR	National Index
Germany	Hypo Real Estate Holding	Quarterly	Transaction prices both cash and mortgage for new dwellings and existing dwellings	Association of German Pfandbrief Banks (vdp): Transactions of 13 financial institutions Expert Committees for Property Valuation GEWOS	Hedonic Regressions	National Index
United Kingdom[16]	The Nationwide Building Society	Monthly	Approved mortgage transactions	Nationwide: Mortgage Data	Hedonic Regressions (strictly cross-sectional hedonic model SCS)	National Index + 13 regions in the UK

[15] Capital city, big cities, small cities, urban area and rural areas.
[16] Wood, R. (2005), A Comparison of UK Residential House Price Indices.

United Kingdom	The Halifax	Monthly	Approved mortgage transactions	Halifax: Mortgage Data	Hedonic Regressions (strictly cross-sectional hedonic model SCS)	National Index + 12 regions in the UK
Ireland	Central Statistics Office (CSO)	Monthly	All transactions	Irish Tax Office	Hedonic Regressions	National Index + Capital (all dwellings)
Finland	Official Statistics of Finland (OSF)	Monthly	All transactions	Finnish Tax Administration's asset transfer tax data	Hedonic Regressions	National Index + categories[17]
France	National Insitute of Statistics	Quarterly	Sample	No direct sources	Weighted average indices adjusted with hedonic methods	National Index + Capital City + Suburbs (all dwellings)
Norway	Statistics Norway	Quarterly	All transactions	FINN.no (website) Real Estate Norway, covering transactions of existing dwellings	Hedonic Regressions	National Index + Capital City + Suburbs (all dwellings)
Poland	National bank of Poland	Quarterly	Sample	380 district governor's offices (local administration units)	Hedonic Regressions	National Index + categories[18]
Spain	National Institute of Statistics	Quarterly	All transactions	Notarial Certification Agency (ANCERT).	Mixed Method: Combine hedonic regressions with stratification	National Index
Repeated Sales						
United States	S&P/ Case-Schiller National Home Price Index	Monthly	Used dwellings transactions	Fiserv Inc[19]	Repeated Sales	National Index + 20 metropolitan regions
Colombia	Central Bank of Colombia	Quarterly	Used dwellings transactions with mortgage	Information on mortgage loans reported by various financial institutions	Repeated Sales	National Index
Canadá	National Bank of Canada / Teranet	Monthly	Used dwellings transactions	Teranet (electronic property registration system)	Repeated Sales	National Index + 11 metropolitan regions

[17] Capital city, big cities, urban area and rural areas.
[18] Capital city, big cities, small cities, urban area and rural areas.
[19] Provider that collects data from multiple sources and then cross-checks them.

United Kingdom	Office for National Statistics	Monthly	Used dwellings transactions	HM Land Registry (Real Transactions)	Repeated Sales	National Index (just for England and Wales)

IPV Methodology

In 2014, the Central Bank of Chile began calculating the IPV based on administrative data from SII, corresponding to actual housing transactions nationwide registered through form F2890, complemented with information from the Real Estate Registry (CBR). In 2014, the method of stratification was defined with 14 strata, constructed through the variables of geographic zone (seven zones) and type of housing (houses and apartments). The calculation requires a series of depurations and imputations that allow for comparable information between strata over time. In issue 107 of the series of Statistical Economic Studies of the Central Bank of Chile, published in June 2014, the methodology and main results of the IPV are presented.

Later, in 2019, the data imputation process and methodology were improved[20] to expand it to 27 strata defined by the variables of geographic zone (seven zones), type of housing (houses and apartments), and condition of the housing (new and used). [21] Finally in 2021, we generate a new set of improvements in the depuration methodology which are explain in the document 139 of the series of Statistical Economic Studies of the Central Bank of Chile, published in July 2023.

I. Database

The IPV uses two sources of information for its calculation, both corresponding to quarterly data provided by the SII:

1- Declaration Form for Real Estate Transfer and Registration (F2890)

Quarterly, the SII provides the BCCh with a database of transactions recorded in the F2890. Additionally, since 2015, the SII has made available to the Central Bank the information from the online F2890 form, which is processed online in the main notaries of the country. Having the online F2890 form reduces the lag time for publishing the IPV, as this information is provided monthly and the lag time for registering sales is less than two months. [22]

2- National Registry of Habitational and Agricultural Real Estate (CBR)

This information is used to complement the information available through the F2890, mainly to impute the variables target code (variable that identifies properties for residential use), built square meters, land square meters and address information (that allows for better classification of houses and apartments) when this information is not available.

[20] Press Release Central Bank of Chile (2019). "Actualización metodológica del Índice de Precios de Vivienda que elabora el Banco Central de Chile".
[21] In the City Center of Santiago, there are no transactions for new houses, so this stratum is removed, resulting in 27 strata instead of 28.
[22] However, the online F2890 is preliminary information, subject to corrections, and does not include all the variables of the regular F2890, and more variables may have missing values.

II. IPV Calculation

The first stage of calculating the IPV involves a series of processes to consolidate the different sources of information and clean the data from these administrative records[23].

These stages can be summarized by the following set of processes:

1- Compilation and formatting of databases: joining the databases, modifying the variables to the required formats, and creating missing variables in the process
2- Removal of duplicates: first, general observations that are identical in all variables, and then a precise specification (ID, date, amount) that refer to the same property in the same period
3- Correction of data entry errors: review and resolution of problems related to data entry
4- Detection of social housing: Elimination of social properties
5- Classification of property type: identification of transactions between houses and apartments
6- Imputation with historical information: destination code variables, built square meters, appraisal, land, and date of registration.
7- Identification of warehouses
8- Imputation of construction

After the data cleaning process, the IPV is calculated. This process can be summarized in the following steps:

1- For each of the 27 strata and for each quarter, the simple average of transaction prices in Unidades de Fomento (currency indexed to inflation) per square meter (UF/m^2) is calculated.
2- A weighted average price is calculated for each quarter. The weights are annual and obtained by dividing the sum of square meters transacted in the stratum during the previous year by the total square meters transacted in the previous year.
3- A chained quarterly price is calculated using prices from the previous year.
4- The house price index is calculated using as a base the chained average price from 2008.

As mentioned in earlier sections, besides the 3-month lag in receiving the information, the IPV stabilizes after two quarters. This is mainly because the registration of new housing units takes longer due to the longer period between sale and delivery of the property. Therefore, the IPV in its last two quarters is considered provisional. For more detail, review the methodological document[24].

[23] The F2890 form is manually entered, which means that this information is not free from errors.

[24] Balsa, J y Vásquez, J (2023). Índice de Precios de Viviendas Banco Central de Chile 2022. Estudios Económicos Estadísticos N°139.

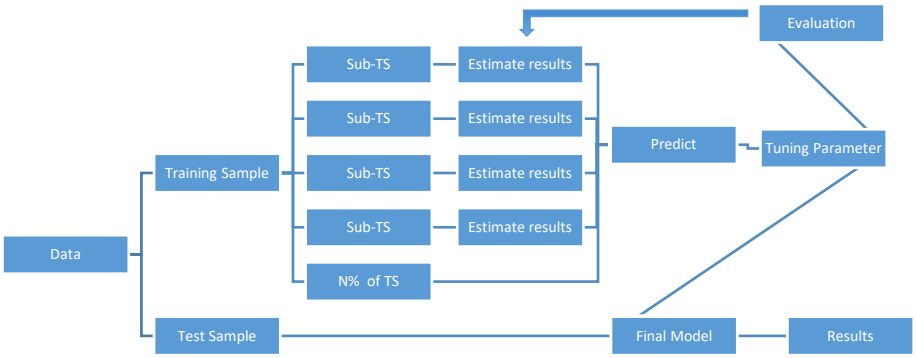
Classification Algorithm

I. What is a Classification algorithm and how it works

Classification algorithms are used to classify data into different classes based on specific given criteria. According to Alpaydin (2010) and Kotsiantis (2007), these algorithms analyze incoming data, learn the patterns and relationships between the various features of the data, and make predictions about the class to which the data belongs. The main goal is to develop a model that manages to predict the class of new data based on its features. The classification process encompasses data preprocessing, feature selection, model training (if need it), evaluation, and finally deployment.

For this paper, we will focus on supervised classification algorithms. As explain in Balsa (2018), Figure I resume the process of a supervised algorithm: first it splits the sample into a training sample and a test sample. Then, for the training sample, the algorithm will create subsamples and leave a n percent apart. After that, the algorithm will start to estimate results from the subsamples and use them to predict the outcomes of the n percent that it was left aside. Later, the tuning parameter is chosen, based on the one that minimizes the loss/cost function, that is normally defined as the sum of the squared residuals, the entropy or other measures in the cross-validation samples. The final model performance is assessed by calculating the lost/cost function of model predictions on the held-out test sample, which was not used at all for model estimation or tuning.

Figure I: Supervised ML Algorithm



There are several types of supervised classification algorithms, including decision trees, kneighbors, random forests, logistic regression, neural networks, among others. Hastie, Tibshirani, and Friedman (2009) and Bishop (2006) note that these different algorithms work in different ways, such as logistic regression that uses a logistic function to model binary dependent variables, decision trees that make decisions by partitioning the data based on feature values, and neural networks that use interconnected nodes to process information and make predictions.

As expected, each algorithm has its own strengths and weaknesses. For example, a logistic regression is easy to interpret, and can handle multi-class classification problems, according

to Kotsiantis (2007). However, it may not work appropriately with non-linear boundaries and can also have high bias. Decision trees, as described by Alpaydin (2010), are easy to interpret and visualize capturing non-linear relations between features. Nevertheless, they can be prone to overfitting, and it is very sensitive to small changes in the data. Lastly, neural networks, as highlighted by Hastie, Tibshirani, and Friedman (2009), are highly flexible and can learn complex decision boundaries, making them very effective for all kind of classification problems, but they can be difficult to interpret, and they consume significant computational resources, making them slower than other algorithms.

One popular classification algorithm is the k-nearest neighbors (k-NN) algorithm, which is a type of instance-based learning. According to Alpaydin (2010) and Bishop (2006), the k-NN algorithm assigns new data points to the class of most of its k-nearest neighbors in the training data set. The value of k is typically chosen based on a validation set, and smaller values of k can lead to higher variance and overfitting, while larger values can lead to higher bias.

The k-NN algorithm has several strengths and weaknesses. As highlighted by Alpaydin (2010), the algorithm is simple and easy to understand, and can be effective for low-dimensional data sets. It can also handle multi-class classification problems and non-linear decision boundaries. However, the algorithm can be computationally intensive and may require significant memory storage, especially for large data sets. In addition, it can be sensitive to outliers and noise in the data.

Overall, the k-NN algorithm is a useful classification algorithm for low-dimensional data sets and can be effective for multi-class classification problems. However, it may not be the best choice for large data sets or when computational resources are limited.

These sources provide a comprehensive overview of classification algorithms and their role in machine learning. They highlight the different types of classification algorithms, how they work, and their strengths and weaknesses. The information provided is useful for anyone interested in understanding the classification process in machine learning.

II. Examples of Property Classification with administrative data

As far as we are concerned, administrative data can be a valuable source for property classification. In this section, we present some examples of countries that use administrative data to classify properties, as well as the methodologies they use.

1. United Kingdom:

The Valuation Office Agency (VOA) uses administrative data from local authorities and the Land Registry for property classification. This data can include information about property location, size, type, age, and architectural characteristics. Although specifics about the use of machine learning are not publicly available, the agency has mentioned in general terms the use of machine learning algorithms for property valuation.

2. Canada:

The Municipal Property Assessment Corporation (MPAC) uses data from property sales, physical inspections, building permits, and other local authority data, using variables that include property size, location, age, type and data from physical inspections. They often use a blend of statistical modeling and machine learning for their methodologies. For example, they use multiple regression analysis to assess property values, considering different variables such as property size, location, among others. As for property classification tasks, such as distinguishing residential properties from commercial ones, they often use decision tree algorithms.

3. Netherlands:

The Dutch Land Registry (Kadaster) and Statistics Netherlands (CBS) use administrative data for property classification. The use a range of this data including property transaction data, physical property attributes. To do this, these organizations use different automated valuation models (AVMs) which could include decision tree algorithms, random forests, or neural networks, trained on historical transaction data.

Overall, administrative data can be useful to get a comprehensive view of the property market, as it includes information on various characteristics of the properties, such as location, size, type, and condition. While there are some differences in the specific variables considered and the algorithms employed, the main goal for organizations is to provide an accurate and fair property valuation. By using administrative data, governments and organizations can create models that assign a value to each property based on its unique characteristics. In addition, these models can help to identify trends and patterns in the housing market over time.

IPV Classification Methodology

As explained earlier, the IPV is obtained using the mixed stratification method, for which three variables are used: property type (house or apartment), condition (new or used), and geographic zone.

The property type variable does not exist in the F2890 database or in the auxiliary databases of the process. However, it is constructed based on the available variables in these sources through an iterative process. Two strategies were followed: one called "Expert Criteria" that uses logical criteria from the researcher to generate the classification, and another strategy called "Data-driven" that uses machine learning algorithms to classify the properties.

Although one might think that it is easy to classify properties using only a criterion based on the square meters of land, where all properties with zero land should be classified as apartments, this is no true. For example, in the case of Type A houses in condominiums (i.e., houses located on common land), even though they are registered with zero land they are houses and have shared land with other houses in the condominium.

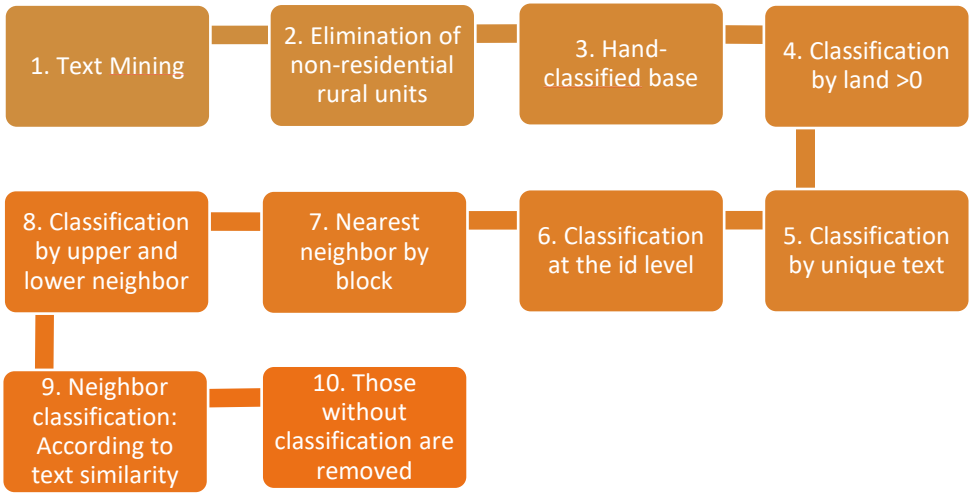
In this section we explain both criteria and the results and uses of the data generated from the process.

I. Flow 1: Expert Criteria

The "expert-criteria" classification process of the IPV works through 10 stages, which are summarized in Figure II:

1. *Text-Mining*: We review the text variables related to the address and look for characteristics that indicate whether they are houses or apartments, as well as whether the properties are rural (plots, land, etc.).
2. *Elimination of non-residential transactions (rural units)*: All properties that were classified as rural in the previous stage are deleted. (3.3% transactions eliminated).
3. *Hand-classified base:* A historical database is created with different units and it is checked with external databases, having around 40.000 units. (It is in continuous growth). (1.27% classified).
4. *Classification by land over zero*: if the Land information is over zero, we assume that it is a house. (58.2% classified)
5. *Classification by unique text:* if was only classified in one type we keep that one. For example, if the address only has information that says is an apartment, we keep it as an apartment. (36.6% classified)
6. *Check classification at the id level:* if all the transactions of a unit have the same classification, we maintain it if not we de-classified the unit. (0.2% classified)
7. *Nearest neighbor by block*: If all the units in a block had the same classification, we complete the ones without classification by the one in the unit. (0.01% classified)
8. *Classification by upper and lower neighbor*: As the name says if my neighbor unit upper and lower has the same classification, we imputed that classification to the unclassified unit. (2.6% classified)
9. *Neighbor classification according to text similarity*: We compare with soundex and levenshtein how similar are two addresses are and we classify by it. (0.1%)
10. *Elimination of non-classify observations*: We eliminated all observations that are not classify. (less than 0.001%)

Figure II: Expert Criteria Flow Resume



The "Expert Criteria" classification process, although works well in correctly classifying around 3 million properties, had errors when classifying new properties. This is because new properties have less information available regarding their addresses, and their land information may not be up to date. For example, Images 1 and 2 show two cases where the "Expert Criteria" misclassified properties as houses when they are apartments.

Image 1: Misclassification in the "Expert Criteria"

Image 2: Misclassification in the "Expert Criteria"

These errors, although they may seem to involve a few units, have a significant impact on the IPV. Since it is a stratified index with 27 zones and a quarterly frequency, the inclusion of hundreds of apartment units (which have a lower price per square meter compared to houses) can result in variations of over 2% in the annual variation of the index in a specific stratum and even up to 1% in the overall index.

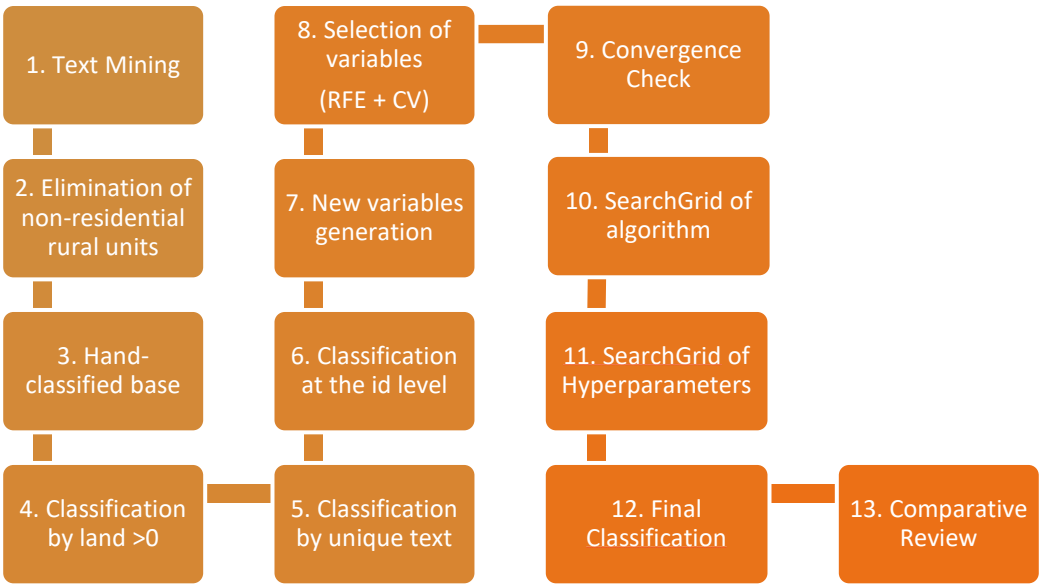
II. Flow 2: Data-Driven classification

Given the errors of the "Expert Criteria", it was decided to experiment with new automated classification methods, which led to a supervised machine learning algorithm.

Like the previous criterion, the first 6 steps were maintained, which were used as the training database for the algorithm. This database consists of approximately 3 million transactions. The selection of the expert criterion cutoff at this step is based on the absence of classification errors up to that point and the consideration of "solid" classification criteria in a complementary manner.

The "Data-Driven" method is explained below (Figure III) with some examples of the code used:

Figure III: Data-Driven Flow Resume



7. *New variables generation:* The database is linked to other sources such as the census and variables are created by block and municipality related to the percentage of each type of property, credit ratio, number of units, among others.
8. *Selection of variables*: We use a recursive feature elimination method with a Random Forest algorithm and five splits of the database for cross-validation. The method chooses ten variables: *Construction*, *Land*, *Credit Ratio*, *Price Ratio*, *Construction Ratio*, *Number of units per block*, *Two measures of quality*, *percentage of apartment by the last census*, *logarithm of the public valuation of the property*.

```
#%% 1.8♦. Feature Selection

"""
Recursive Feature Elimination with cross validation
"""
min_features_to_select = 1  # Minimum number of features to consider
rf = se.RandomForestClassifier(random_state=123)

rfecv = RFECV(
    estimator=rf,
    step=1,
    cv=cv,
    scoring="accuracy",
    min_features_to_select=min_features_to_select,
    n_jobs=25,
)
rfecv.fit(x, y)

print(f"Optimal number of features: {rfecv.n_features_}")

# get the selected features
selected_features = np.array(x.columns)[rfecv.support_]
print("Selected features: ", selected_features)
```

9. *Convergence Check*: One important step is to verify if all the chosen algorithms converge to results. Without this preliminary step, the code could run indefinitely without reaching a conclusive outcome. SVM was eliminated for non-convergence.

```
#%% 1.9. Model Selection
"""
Convergence Check
"""
rf   =   se.RandomForestClassifier(n_estimators=100,   max_depth=5,   n_jobs=6,
random_state=123)
rf.fit(x_train, y_train)
rf_pred=rf.predict(x_test)

svm = SVC(C=1, kernel="rbf", gamma="scale",  random_state=123)
svm.fit(x_train, y_train)
svm_pred=svm.predict(x_test)

lr = LogisticRegression(C=1, solver="lbfgs",  random_state=123)
lr.fit(x_train, y_train)
lr_pred=lr.predict(x_test)

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(x_train, y_train)
knn_pred=knn.predict(x_test)
```

10. *Searchgrid of algorithms*: We use the search grid technique to systematically explore different combinations of algorithms and hyperparameters to find the best model performance. This helps automate the process of finding optimal hyperparameter settings.

```
#%% 1.9. Model Selection
"""
Model SearchGrid
"""
x= x[Ganadoras]

classifiers = [
    se.RandomForestClassifier(n_jobs=7,random_state=123),
    LogisticRegression(n_jobs=7),
    KNeighborsClassifier(n_jobs=7),
    ]
param_grids = [
    {
        "n_estimators": [100,200,300],
        "max_depth": [3,5,7],
     },
    {
     "C": [0.1,1,10],
     "penalty": ["l1", "l2"],
     "solver":["lbfgs","liblinear"],
     },
    {
     "n_neighbors": [3,5,7],
     "weights": ["uniform", "distance"],
     "p":[1,2],
     },
    ]

best_score=0
best_params= None
best_model= None
```

```
counter=0

for classifier, param_grid in zip(classifiers, param_grids):
    print(fr"starting {counter} of 255")
    grid_search= GridSearchCV(classifier, param_grid, cv=5, n_jobs=7)
    grid_search.fit(x_train, y_train)

    if grid_search.best_score_ > best_score:
        best_score= grid_search.best_score_
        best_params = grid_search.best_params_
        best_model = grid_search.best_estimator_
    counter=counter+1

test_score=best_model.score(x_test, y_test)
print("best Score:", best_score)
print("best Parameters:", best_params)
print("Best model:", best_model)
print("Test Score:", test_score)
```

11. *Searchgrid of Hyperparameters*: We use the same technique to choose what hyperparameters give the best model performance for the algorithm that won. [25]

```
#%% 1.9. Model Selection
"""
Hyperparameter searchgrid
"""
classifiers = [
    KNeighborsClassifier(n_jobs=10),
    ]
param_grids = [
    {
     "n_neighbors": list(range(1,15)),
     "weights": ["uniform", "distance"],
     "p":[1,2],
     "leaf_size": list(range(1,30)),
     },
    ]

best_score2=0
best_params2= None
best_model2= None

cv = KFold(n_splits=5)

for classifier, param_grid in zip(classifiers, param_grids):
    grid_search=   GridSearchCV(classifier,   param_grid,   cv=cv,   n_jobs=3,
pre_dispatch='n_jobs')
    grid_search= grid_search.fit(x_train, y_train)


    if grid_search.best_score_ > best_score2:
        best_score2= grid_search.best_score_
        best_params2 = grid_search.best_params_
        best_model2 = grid_search.best_estimator_

test_score=best_model2.score(x_test, y_test)
```

[25] Care must be taken with these steps since the excessive search of hyperparameters in very small databases can lead to overfitting. Therefore, all processes are run with a cross-validation process to avoid sample bias.

12. *Final Classification*: The transactions are classified with the selected algorithm and hyperparameters.
13. *Comparative Review*: We compared the results of Flow 1 with Flow 2 and manually reviewed the differences.

All these previous processes were run in Python 3.7 using the modules of Sci-kit learn, numpy, pandas, joblib and os. The database had 34 variables and a training set of 3M transactions with 21 variables pre-selected with expert criteria. The algorithm selects 10 variables: *Construction*, *Land*, *Credit ratio*, *Price ratio*, *Construction ratio*, *N° of unit per block*, *Quality 1 and 2*, *percentage of apartment by Census*, *Log of public valuation* and runs with a cross-validation of 5 splits.

We compare the results of 4 algorithms: *Random Forest*, *Support Vector Machine*, *Logit*, *Kneighbors*, and we choose the algorithm with best accuracy. The winner was Kneighbors with the following hyperparameters:

Neighbors = 13 | weights = 'distance' | p = 1 | leaf size = 15

The algorithm predicts with an accuracy of 99.2% in-sample and 98.9% out-of-sample. When comparing the classification of flow 1 with flow 2, we can see that flow 2 correctly classifies the buildings solving the previous problems, but it does have some issues with new houses. That's why we decided to use both processes and compare them.

It's also important to note that the process improves with each iteration through the hand-classified database. The first time we ran it, the difference between Flow 1 and Flow 2 was 250,000 transactions, and it has been decreasing to 7,000 transactions.

This process is also re-tunable and is recommended as market trends change or updates to external data sources are incorporated.

III. Bonus Track: ML Algorithm for any classification problem

During the creation of this classification process a change of emphasis was made from pure property classification to an easy-to-use code for any classification process of the Central Bank of Chile.

This process can be used for both dichotomous and multicategory problems, as well as focusing on specific problems such as unbalanced samples, all through the choice of algorithms to incorporate in the searchgrid or limiting the searchgrid within a subset of hyperparameters. For example, if the problem is highly unbalanced, one can choose Kernel-Support Vector Machine or Logit with an L1 regularization.

In order to be able to modify the code, it is only necessary to have a pre-classified database to train the algorithms. However, it is also advisable to have some criteria for the pre-selection of variables.

IV. Results and uses of the data

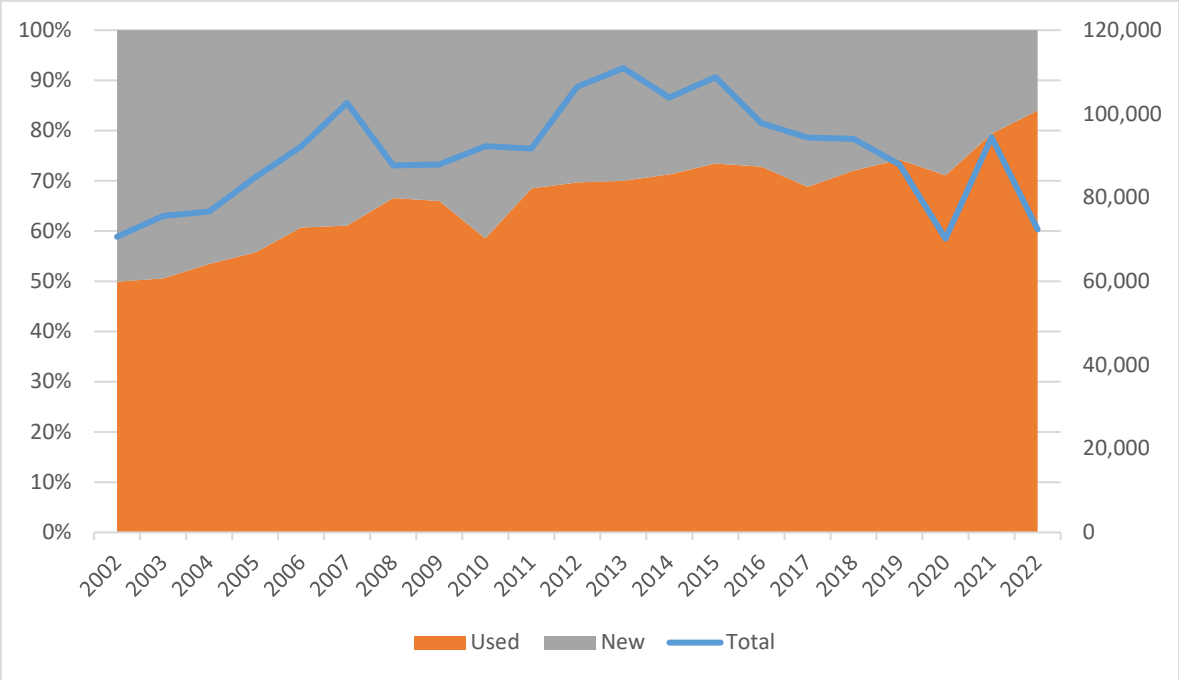
The purpose of this section is to show how the Central Bank uses the data classified by the algorithm. The final consolidated and cleaned database contains data from the first quarter of 2002 through the fourth quarter of 2022 and has approximately 3.2 million residential property transactions nationwide. Of these transactions, 47.8% correspond to properties in the Metropolitan area, 61.2% to houses and 42.4% to new properties. The consolidated and cleaned database has an average price of 27.4 UF/m2 and a standard deviation of 19.33 UF/m2, the maximum price value in UF/m2 is 200.3 UF/m2.

The number of annual transactions increased by 35% between 2002 and 2022[26] (76% in 2021), from 94 thousand to over 125 thousand. However, since the peak in 2015 (186 thousand), the number of transactions has been declining reaching even less than 130 thousand by 2022. The market for houses (70 thousand) in 2002 more than tripled the number of apartments (24 thousand). This dynamic has been adjusting over the years due to the increase in urbanization and gentrification of cities, with the housing market (72 thousand) being only 30% higher than the apartment market (56 thousand) by 2022.

On the other hand, the apartment market has grown substantially between 2002 and 2022, going from 24 thousand to 56 thousand units annually, finding its peak in 2019 with 82 thousand transactions. The most significant growth is found in new apartments, which from 11 thousand units in 2002 reach 28 thousand in 2021, almost 2.6 times more than in 2002 and reaching up to 4.5 times more in 2017 (50 thousand). Transactions of used apartments increase from 13 thousand to 27 thousand in this same period, with 2021 being the year with the highest number of transactions (38 thousand).[27]

[26] It is important to note that the 2022 base does not yet contain all transactions and could increase by up to 10% based on analyses performed for other years.
[27] Graph N°2.

Graph N°1. Number of transactions per year and composition between new and used (Houses)



Graph N°2. Number of transactions per year and composition between new and used (Apartments)

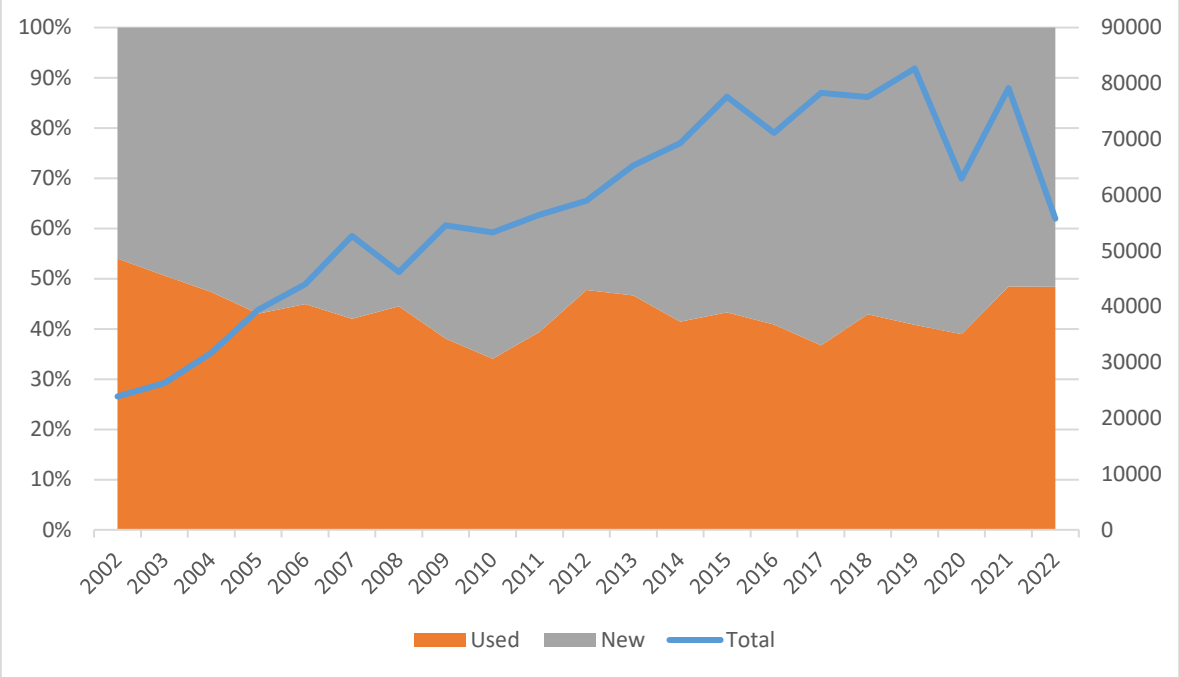


Table N°2 shows the number of transactions during 2022 for each of the 27 strata. In total there were 128 thousand transactions, which are broken down between 88 thousand used

and 40 thousand new, of which 55 thousand corresponded to apartments and 72 thousand to houses. In the Metropolitan Region, the east and west zones registered the highest number of transactions, with 15,803 and 15,348 respectively, with apartments being the most traded in the east zone and houses the most traded in the west zone of the capital. At the national level, the central zone registered some 37 thousand transactions, dominated by used houses (19 thousand).

Table N°2. Number of transactions per stratum (2022)

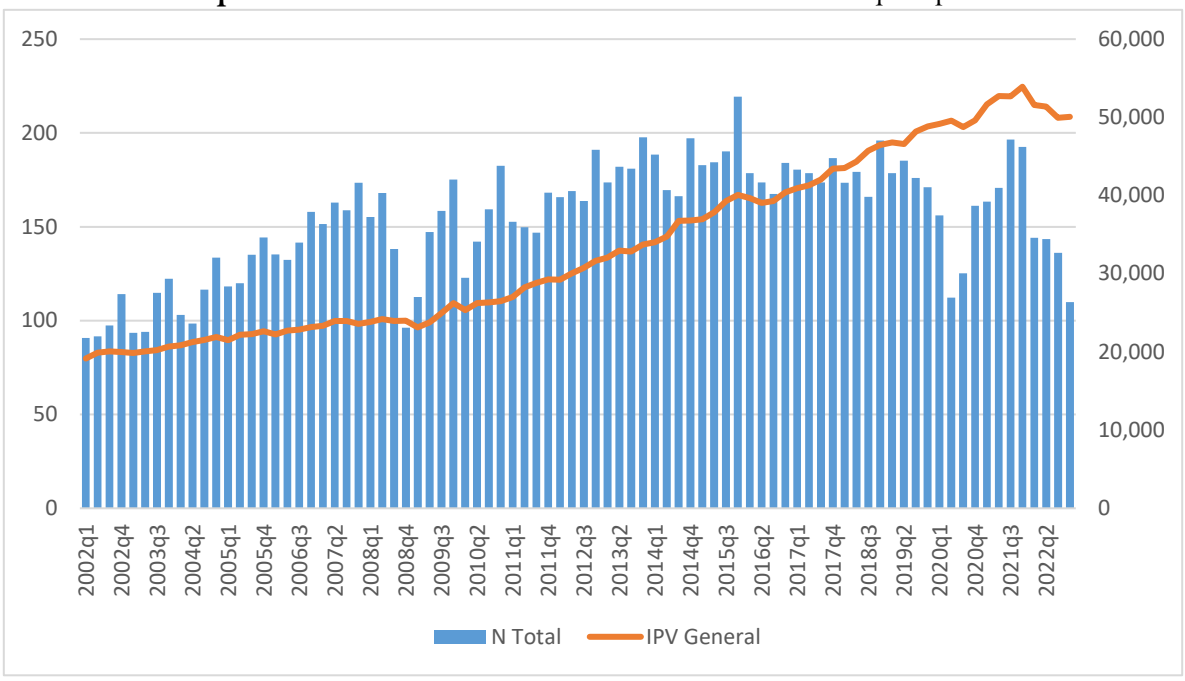
	Houses			Apartments		
Zone	Used	New	Total	Used	New	Total
MR Central	630	0	630	3729	3317	7046
MR East	3935	205	4140	6739	4924	11663
MR West	8587	1777	10364	2541	2443	4984
MR South	7002	1686	8688	2199	3219	5418
Total MR	**20,154**	**3,668**	**23,822**	**15,208**	**13,903**	**29,111**
North Zone	4774	314	5088	2027	1392	3419
Central Zone	18958	4016	22974	6826	7304	14130
South Zone	16878	3594	20472	2909	6171	9080
Total	**60,764**	**11,592**	**72,356**	**26,970**	**28,770**	**55,740**

Graph N°3 and N°4 show the general dynamics between the IPV and the number of transactions. With them, the Central Bank can study the dynamics of this indicator at different junctures:

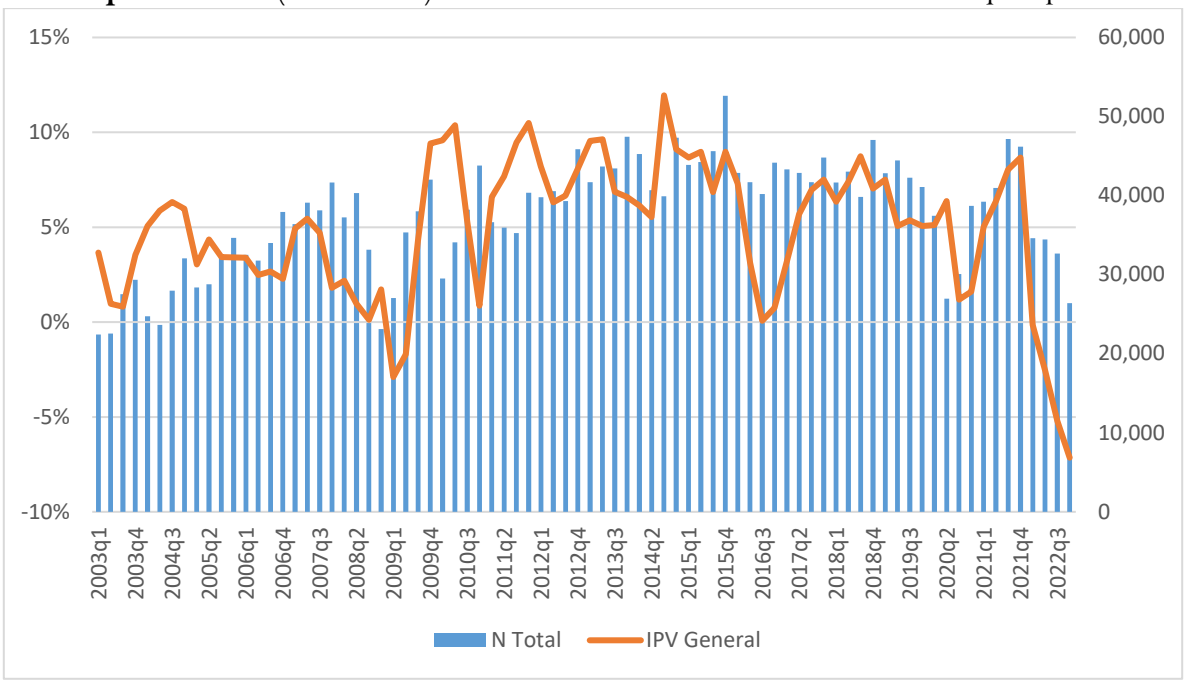
1. Housing prices have risen 161% nationwide from the first quarter of 2002 to the fourth quarter of 2022, with the western part of the metropolitan region being the area with the highest increase (245.3%). In the case of house condition, new homes (175%) are leading the increases compared to used homes (156.8%).
2. In the first quarter of 2009, housing prices had one of their largest declines nationwide (-2.9%), caused mainly by the effects of the subprime crisis. This drop was driven by used homes (-8.6%), particularly houses (-10%), while new homes slowed, but did not enter to negative territory (2.2%).
3. The fall in prices at the end of 2010 is consistent with the effects of the earthquake, which, unlike in the subprime crisis, it was new homes that led the falls (-4%).
4. The increase in prices in 2014 that was linked to a drop in sales of construction companies explained by the economic conditions of the time (11.9%).
5. The confluence of factors derived from the social outbreak and the pandemic, evidenced by an increase in credit restrictions and the unemployment rate, as well as the decrease in real remunerations, have caused a notorious repercussion in the number of transactions and in the decrease of the price by 11.9% as of the fourth quarter of 2022. This drop is the largest recorded since the existence of the index.
6. Finally, and linked to the previous point, there is a decoupling in the trend between new and used homes, where new homes have remained in positive values during 2022, unlike used homes that have entered negative territory, with the MR (-11.9%)

and especially the eastern zone (-12.6%) contributing the most to the fall in the index.

Graph N°3. General IPV and Number of transactions per quarter



Graph N°4. Var (12 months) General IPV and Number of transactions per quarter



Conclusion

The global housing indexes are of great relevance for the decision making of various agents of the economy, which has increased since the subprime crisis and in the face of the recent pandemic. Their interconnection with the financial system has led the Central Bank of Chile to pay particular attention to them in recent years.

However, the creation of these indicators, especially if they use administrative data, is not without challenges. This paper presented a solution to the handling of this data for the calculation of the Housing Price Index (IPV), managing to classify between house and apartment 3.2 million transactions automatically, with an error rate close to zero and thus reducing substantially the working hours involved (from 5 working days to just one morning).

Interestingly, the process is available not only for property classification, but also as a first approach to any classification problem, whether it is dichotomously balanced, multi-categorical or unbalanced.

Among the main results we found that the IPV constructed with this methodology has had the theoretically expected behavior in the face of the shocks experienced by the Chilean economy and especially by the construction industry.

Particularly, the index correctly reflects the price dynamics caused by the social outbreak and the COVID-19 pandemic, showing a strong fall in housing prices (-11.9%), which is congruent with the greater credit restrictions, the increase in unemployment, the fall in real wages and the deceleration in the construction sector.

It would be interesting to review the results of the methodology in the face of other classification problems, looking at its consistency and effectiveness. It would also be interesting to analyze how much training data is needed for it to work properly.

References

- Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA: MIT Press.
- Balsa, J y Vásquez, J (2023). Índice de Precios de Viviendas Banco Central de Chile 2022. Estudios Económicos Estadísticos N°139.
- Balsa, J. (2018). Using causal tree algorithms with difference in difference methodology: a way to have causal inference in machine learning. Available in https://repositorio.uchile.cl/handle/2250/168527
- Central Bank of Chile (2014). Índice de Precios de Viviendas en Chile: Metodología y Resultados. Estudios Económicos Estadísticos N°107.
- Bishop, C. M. (2006). Pattern recognition and machine learning (1st ed.). New York, NY: Springer.
- Case, K. E., Quigley, J. M., & Shiller, R. J. (2013). Wealth effects revisited 1975-2012. Cowles Foundation Discussion Paper No. 1933.
- Englund, P., Hwang, M., & Quigley, J. M. (2002). Hedging housing risk. Journal of Housing Economics, 11(4), 310-342
- Eurostat (2013). Handbook on Residential Property Prices Indices (RPPIs). Eurostat Methodologies & Working Papers.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.
- Himmelberg, C., Mayer, C., & Sinai, T. (2005). Assessing high house prices: Bubbles, fundamentals, and misperceptions. Journal of Economic Perspectives, 19(4), 67-92.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.
- Owusu-Ansah, A. (2018). Construction and Application of Property Price Indices. Routledge Studies in International Real State.
- Prasad, N. L., y Richards A. (2008). Improving Median Housing Price Indexes Through Stratification. Journal of Real Estate Research, vol. 30, No. 1, págs. 45–71.
- Press Release Central Bank of Chile (2019). "Actualización metodológica del Índice de Precios de Vivienda que elabora el Banco Central de Chile": https://www.bcentral.cl/c/document_library/get_file?uuid=e8d65791-b51b-43a3-3e6b-94fbb52732de&groupId=33528
- S&P Dow Jones Indices. (2021). S&P/Case-Shiller U.S. National Home Price Index. Retrieved from https://www.spglobal.com/spdji/en/indices/real-estate/sp-case-shiller-us-national-home-price-index/

- Wood, R. (2005), A Comparison of UK Residential House Price Indices. Real Estate Indicators and Financial Stability, BIS Papers No 21, Banco de Pagos Internacionales.

Property classification with administrative data: The case of the Chilean House Price Index

Department of Microdata

Statistics Division , July 2023

The Challenge

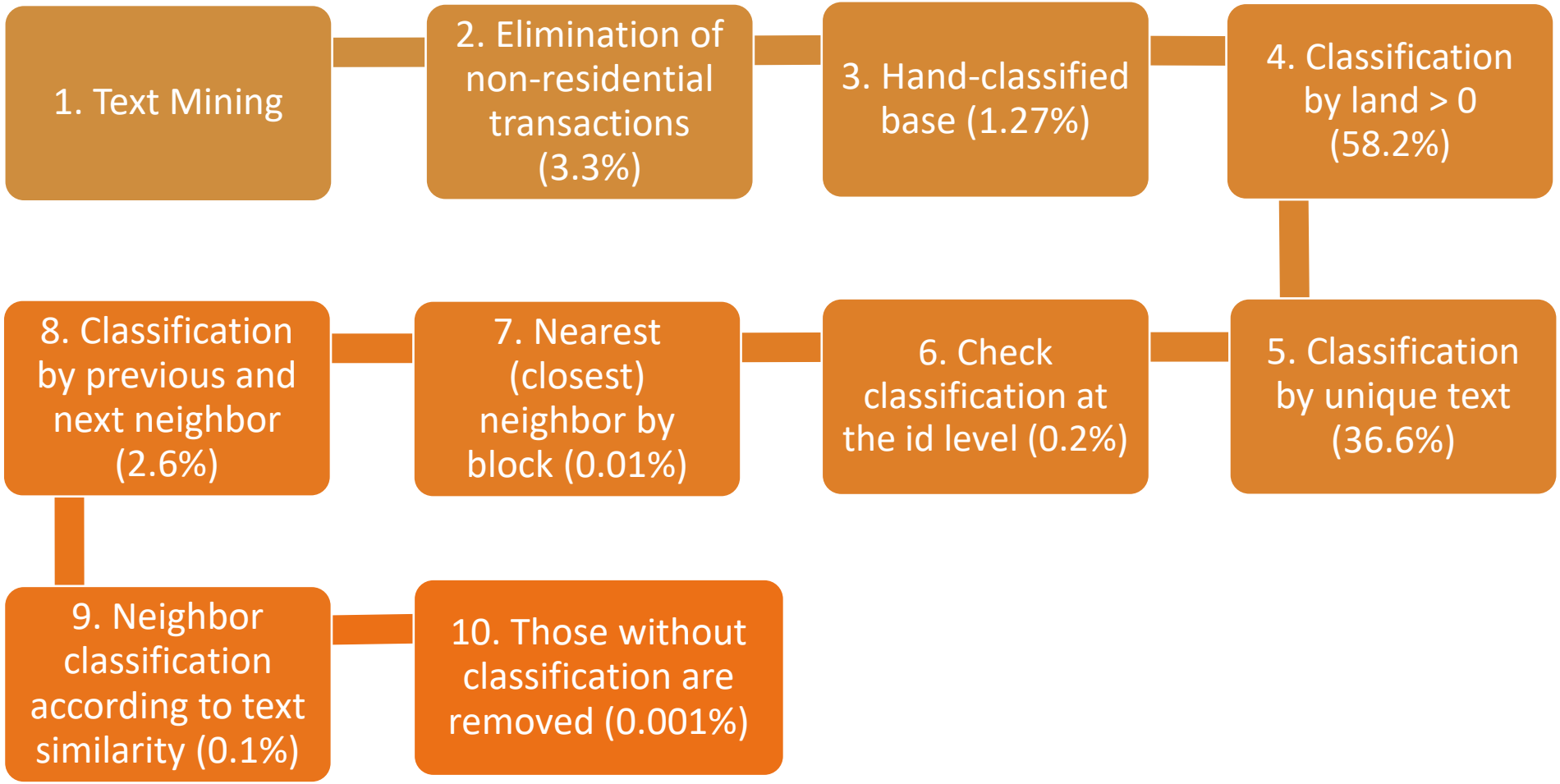
The Chilean House Price Index (IPV)

- Since 2014, the Central Bank of Chile calculates the House price Index based on administrative data from the Tax Administrative office, corresponding to actual housing transactions nationwide registered, complemented with information from the Real Estate Registry.
- The IPV is an experimental statistics and it's created based on a stratification method with 27 strata defined by:
  - Geographic zones (7)
  - Type of Housing (houses and apartments)
  - Condition of the property (new and use
- Of course, the distinction between houses and flats (apartments) is not part of the information reported by the tax authorities, so it has to be classified within the IPV calculation process.

banco central Chile

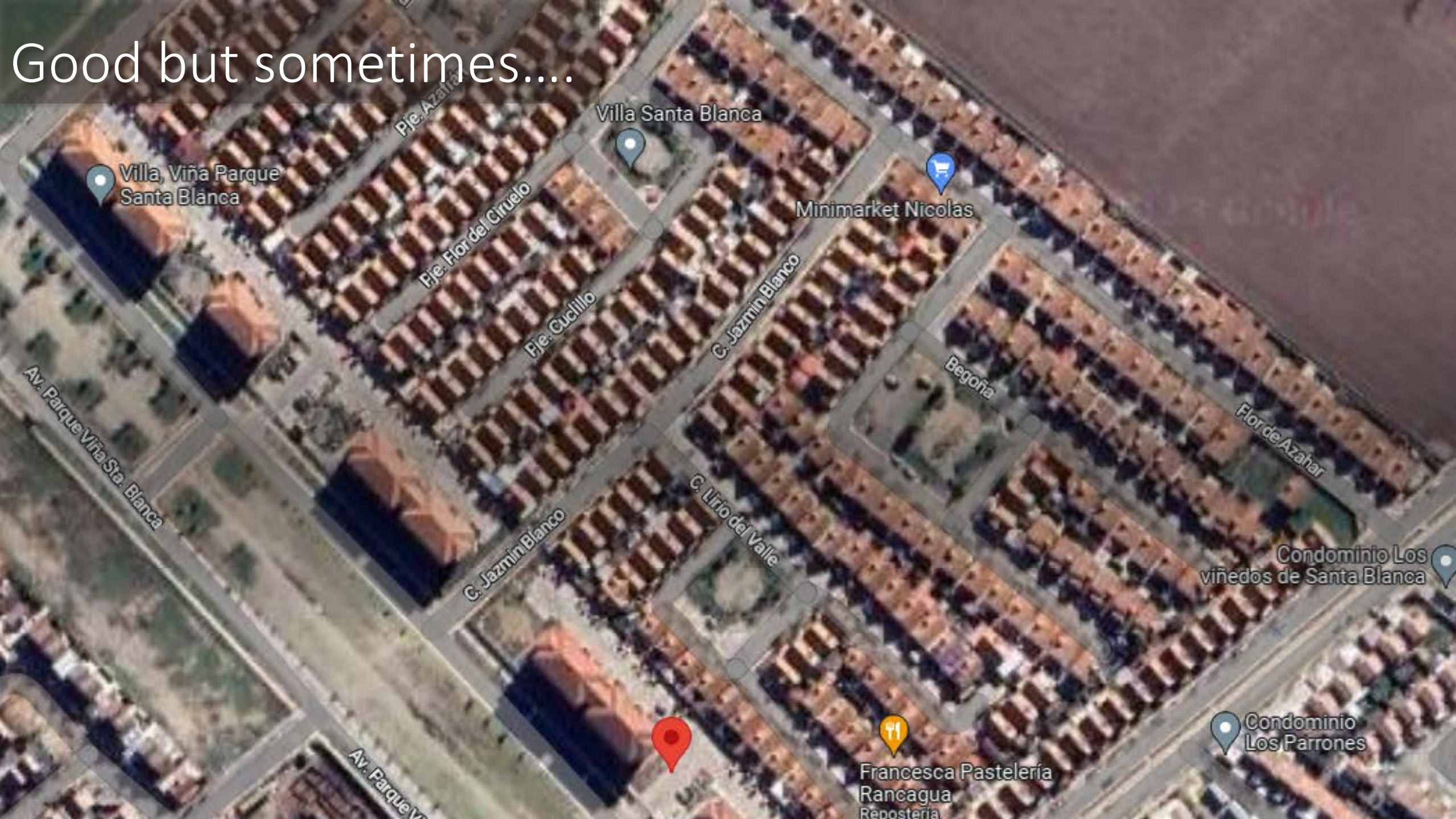
Solution

Manual Process

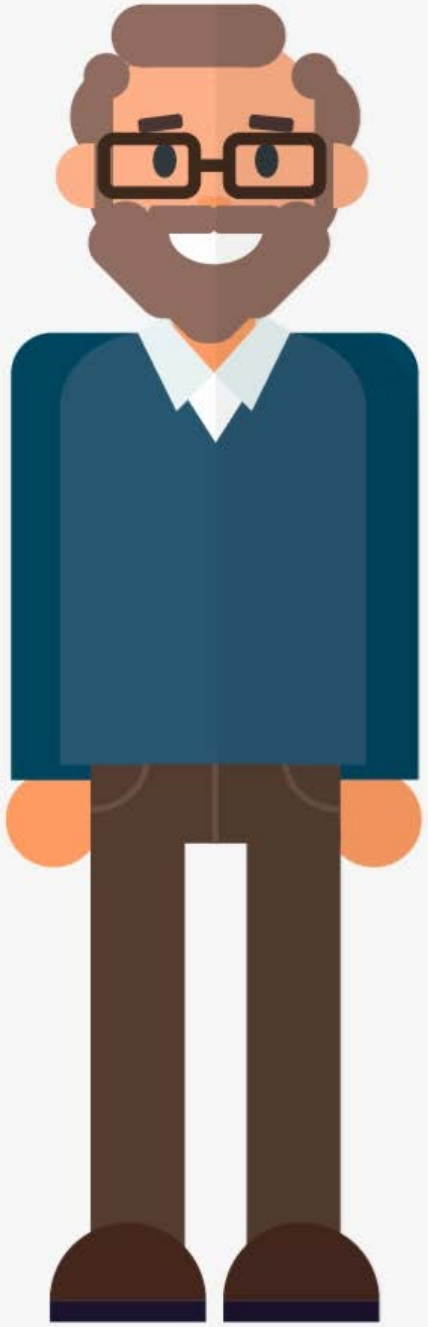


banco central Chile

Good but sometimes....

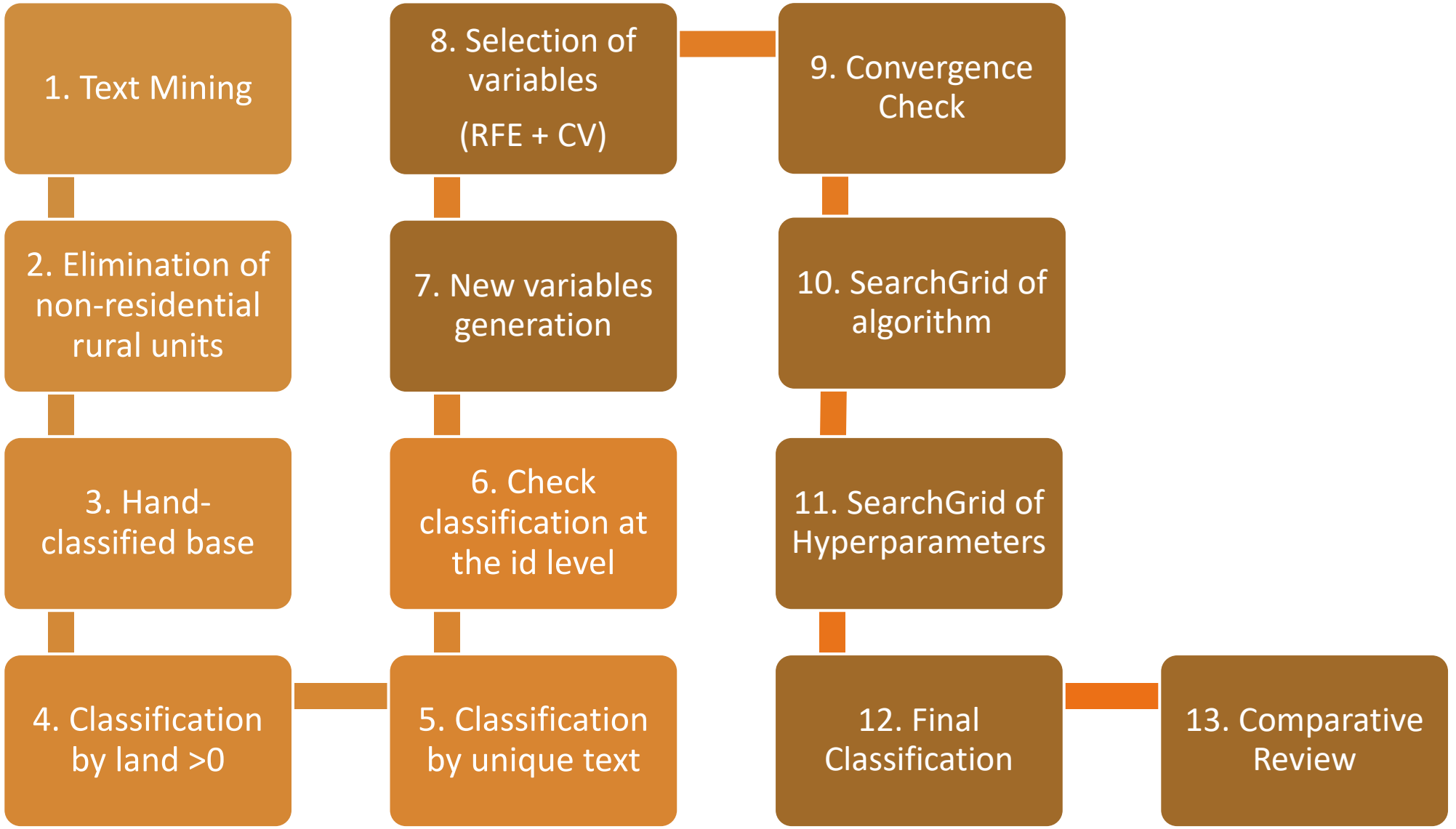


Leave it the way it is? Change or Fix the previous algorithm? Machine Learning - AI?



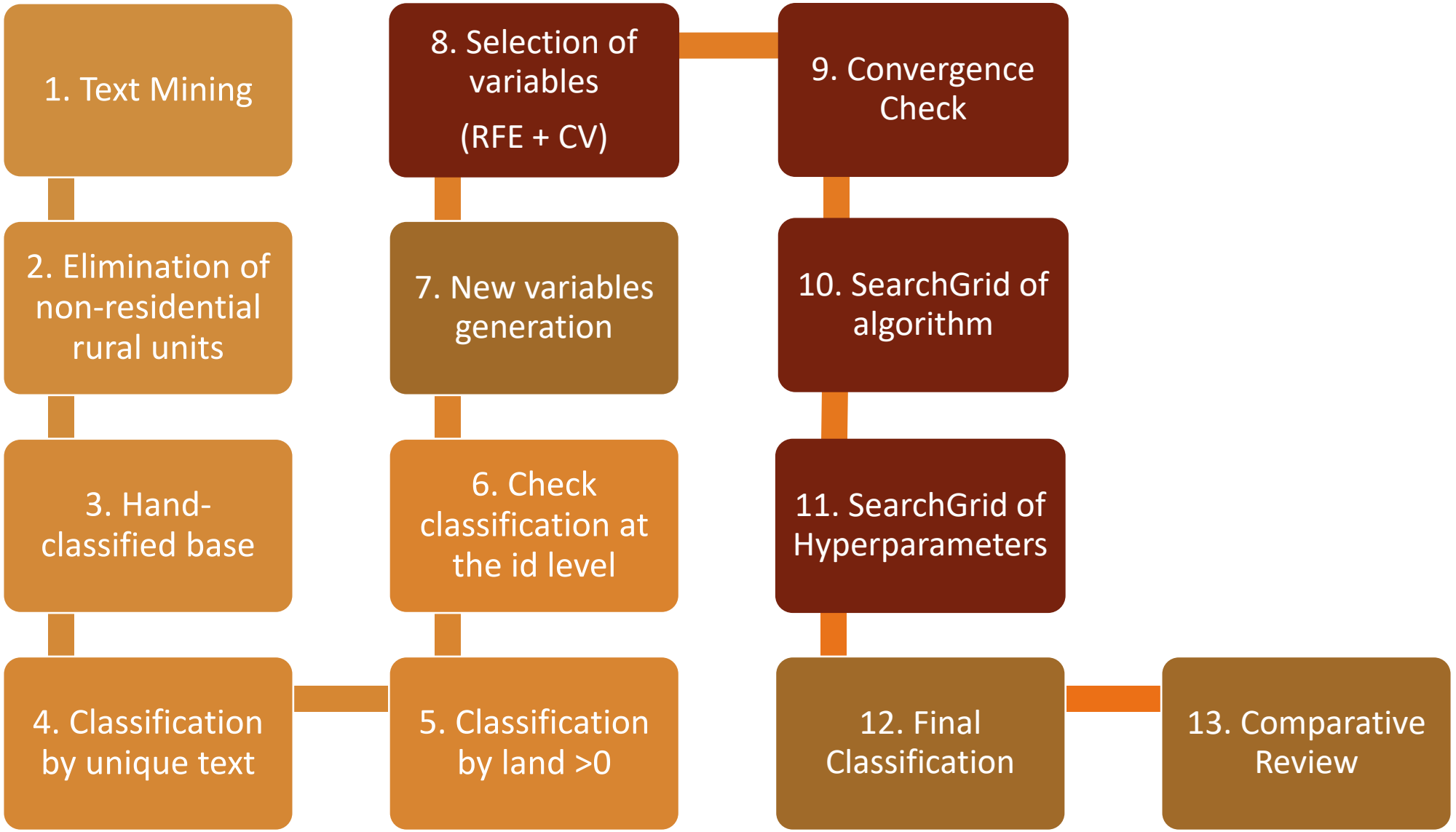
Improved process with ML

Data-Driven Process



banco central Chile

Data-Driven Process



banco central Chile

Characteristics of the Process

- The database had 34 variables and a training set of 3M transactions with 21 variables pre-selected with expert criteria.
- The algorithm selects 10 variables:

  Construction | Land | Credit ratio | Price ratio | Construction ratio | N° of unit per block

  Quality 1 and 2 | % of apartment by Census | Log of public valuation
- The algorithms (CV=5):

  Random Forest | Support Vector Machine | Logit | Kneighbors
- The winner = Kneighbors

  Neighbors = 13 | weights = ‘distance’ | p = 1 | leaf size = 15

banco central Chile

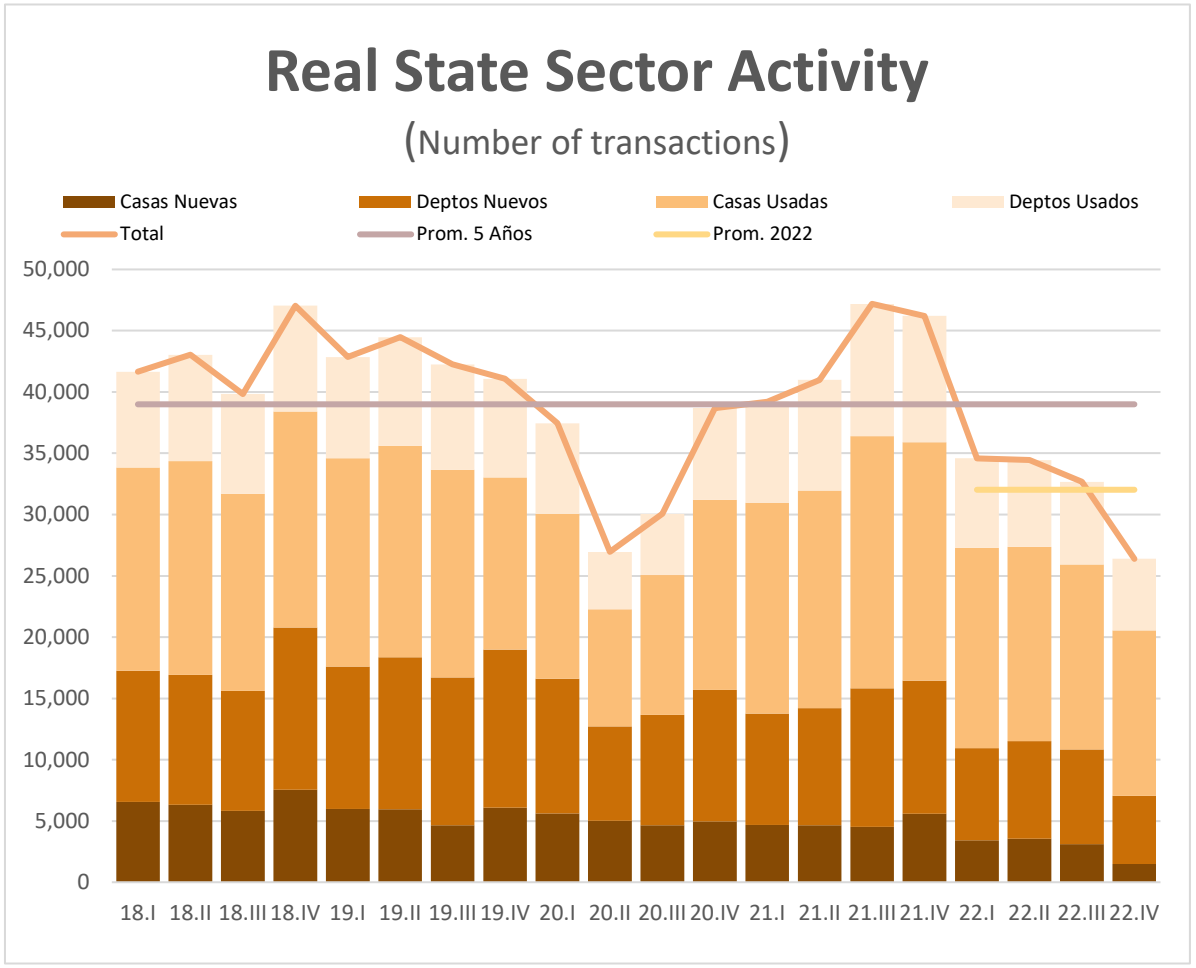
Results

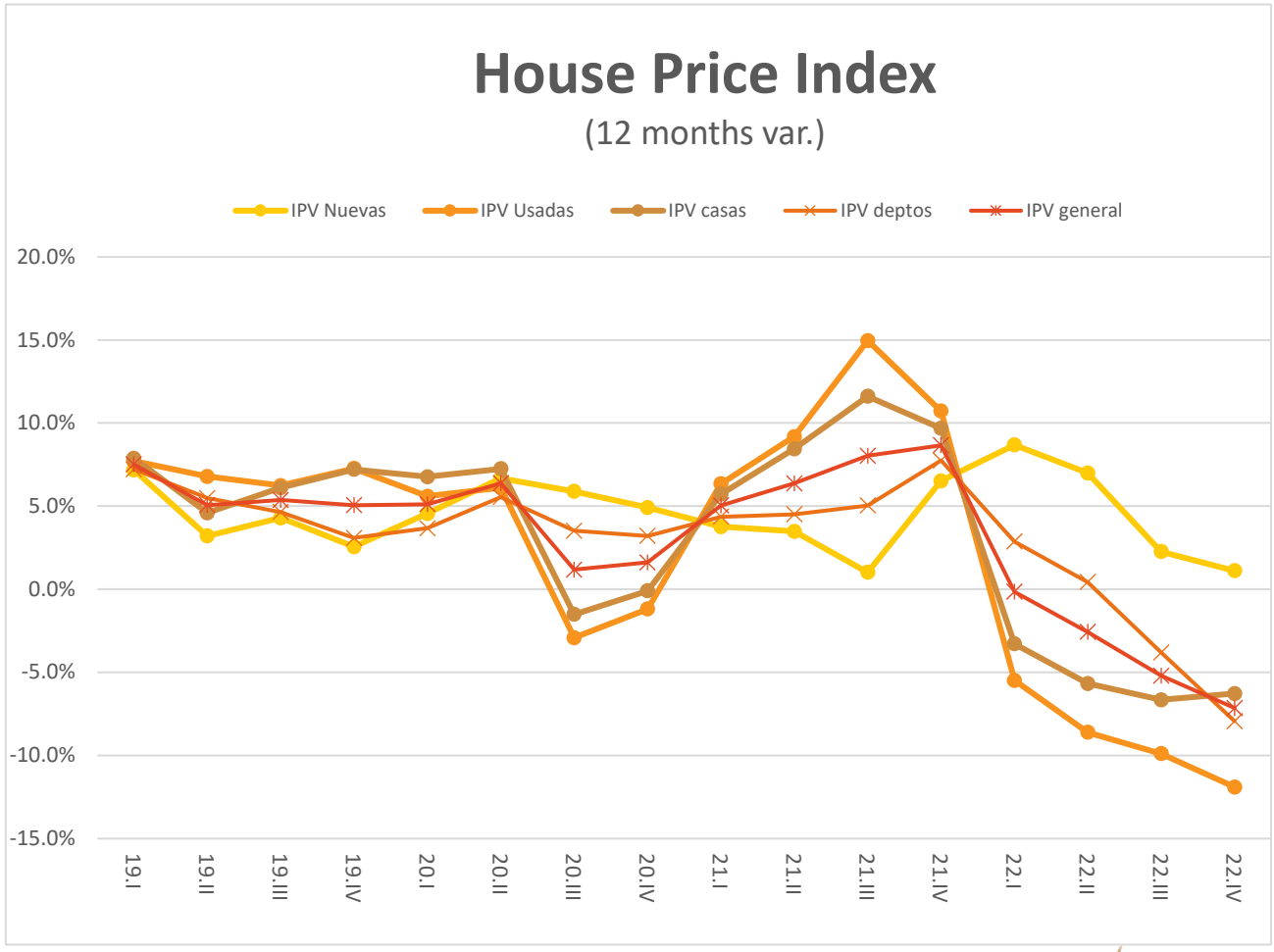
Bottom Line

- The improved process is parallelized and takes 30 minutes
- The process reduces the working hours from several days to one morning with a level of accuracy very close to 0.
- The algorithm predicts with an accuracy of 99.2% in-sample and 98.9% out-of-sample
- The process improve with each iteration through the hand-classified database (250.000 vs 7.000 dif)

banco central Chile

Uses of the data





banco central Chile

Conclusion

Available toolbox for everyone

- Simple process that only requires a training dataset and preferably some expert criteria for pre-selection of variables.
- Process usable for any type of binary problem (extendable to multicategory problems).
- Evolutionary and re-tunable process.

If it is possible don't focus on the specific problem, focus on how to solve the entire family of problems, and share your findings.

Property classification with administrative data: The case of the Chilean House Price Index

Department of Microdata

Statistics Division , July 2023

Annex 1: Stratification or "Composition Adjustment"

What is the stratification method?

- The stratification method, also known as "composition adjustment," is based on the creation of strata or groups of transactions that aim to neutralize variations in the composition of properties sold.
- Stratification involves breaking down the total sample of transactions into a series of sub-samples or strata. Then, for each stratum, the price index based on the mean or median is obtained to obtain the aggregated housing price index as a weighted average of the indices of each stratum. That is:

$$P^{0t} = \sum_{m=1}^{M} \omega_m^0 P_m^{0t}$$

- Where P_m^{0t} is the index of the stratum m that compares the average price of the period t is the period with base 0, and ω_m^0 denotes the weight of the stratum m.
- The formation of the strata is based on certain variables, such as geographic location, type of housing, housing condition, property size, etc. In this way, changes in the composition of housing between different periods are controlled, but not within each stratum.

banco central Chile

Thank you

Juan José Balsa F.
jbalsa@bcentral.cl