mathematics

MDPI

Article

Variable Selection for Meaningful Clustering of Multitopic Territorial Data

Xavier Angerri * and Karina Gibert *

Intelligent Data Science and Artificial Intelligence Research Center and Institut de Ciència i Tecnologia de la Sostenibilitat, Universitat Politècnica de Catalunya-BarcelonaTech, 08034 Barcelona, Spain
* Correspondence: xavier.angerri@upc.edu (X.A.); karina.gibert@upc.edu (K.G.)

Abstract: This paper proposes a new methodology to improve territorial cohesion in clustering processes where many variables from different topics are considered. Clustering techniques provide added value to identify typologies, but there are still unsolved challenges when data contain an unbalanced number of variables from different topics. The territorial feature selection method (TFSM) is presented as a method to select the representative variable of each topic such that the interpretability of resulting clusters is preserved and the geographical cohesion is improved with respect to classical approaches. This paper also introduces the thermometer as a new knowledge acquisition tool that allows experts to transfer semantics to the data mining process. TFSM proposes the index of potential explainability (E_k) as the criteria to select the most promising variables for clustering. E_k is based on the combination of inferential testing and metrics such as support. The proposal is applied with the INSESS-COVID19 database, where territorial groups of vulnerable populations were found. A set of 195 variables with 21 unbalanced thematic blocks is used to compare the results with a traditional multiview clustering analysis with promising results from both the geographical and the thematic point of view and the capacity to support further decision making.

Keywords: data science; intelligent decision support; COVID-19; traffic light panels; thermometer; feature selection; explainable AI; maps; Catalonia

MSC: 68T05; 68T30; 97K80; 90B50

Citation: Angerri, X.; Gibert, K. Variable Selection for Meaningful Clustering of Multitopic Territorial Data. *Mathematics* **2023**, *11*, 2863. https://doi.org/10.3390/math11132863

Academic Editor: Manuel Alberto M. Ferreira

Received: 9 May 2023
Revised: 18 June 2023
Accepted: 21 June 2023
Published: 26 June 2023

1. Introduction

Making global territorial decisions might not be a good option when the territory involves heterogeneity. Very often, several parts of the population behave differently from others. Making sustainable decisions quite often means making different decisions associated with the different scenarios occurring in the different areas of the territory and associating these decisions to smaller areas where individuals are sufficiently similar. However, designing policies for local territorial units might be inconvenient because statistical secrecy might be in danger when the affected groups become too small.

Finding mechanisms to identify groups of similar territorial units that can share the same policies is a good option to support policy making at an intermediate level of territorial granularity that preserves statistical secrecy and goes further than global territory's more general and imprecise decisions. To find a suitable grouping of territorial units, clustering processes oriented to identify blocks of similar individuals to receive the same treatment are indicated. The resulting clusters will contain groups of territorial units sharing a single decision per group.

However, applying clustering directly to datasets of individuals does not ensure that the results are consistent from a territorial point of view. This means that clusters might be formed with similar individuals coming from very different locations in the territory, and this will make it impossible to design policies from a territorial perspective. This was one of the main challenges of that paper and where difficulties have been met.

This paper presents a methodology able to find clusters with territorial coherence from individual databases (microdata).

The proposal also addresses a second challenge that is also common in the field of clustering citizen data, which is dealing with data about different topics that are often represented in an unbalanced way. Indeed, when clusters have to be discovered on a database containing information from different topics and the number of variables from each of the topics is unbalanced, classical clustering is biased towards more represented topics (those with more variables). Multiview clustering is an approach often used to mitigate these biases. However, territorial coherence is not taken into account by the original multiview approach. Therefore, a modification of the multiview clustering is proposed based on selecting representative variables per each topic, thus maintaining territorial coherence.

Thus, the main goal of this paper is to define a new methodology for finding territorial clusters from individual datasets with territorial coherence and the balanced impact of all topics tackled in the data. The resulting clusters will guarantee that all locations in a cluster are close and similar. The presented proposal gives a central role to the location variable. In the general case, the territorial variable describes a certain level of administrative division of the territory (cities, regions, etc.). Our solution is the territorial feature selection methodology (TFSM).

The following are the specific subgoals addressed in this paper with the aim of achieving our overall goal:

- We introduce the semantics of the variables through a new formalism because there is a need to drive the meaning of the variables to the interpretation of results through the analysis process. Our solution is the thermometer.
- We design a new automatic methodology of building a conceptual interpretation of profiles based on the traffic light panel (presented in Section 2.4) method of automatization, which helps the final user understand the results because there is a need for automatically building TLPs, including the semantics of the variables. Our solution is the creation of a TLP based on the thermometer.
- We create new data-driven variables to sum up variables from the same topic because there is the need to balance the number of variables per thematic block for the final global analysis. Our solution is the data-driven 2nd generation indicator.
- We discover which variables should be included in the global clustering to guarantee clusters with territorial coherence because there is a need for objective criteria to choose the most appropriate variable from each thematic block. Our solution is the index of potential explainability.

Accordingly, the main contributions of this paper are:

- The territorial feature selection method: This methodology is used to discover which clusters have territorial coherence. This method is the main contribution proposed in this paper and includes the identification of the variable with more significance in a set of territories.
- The thermometer: This is a new tool that assigns basic traffic light colors (green, yellow, or red) to ranges of values of the numerical variables or to the modalities of qualitative variables so that colors are associated with the semantics of the variable. It is a knowledge acquisition tool that allows domain experts to transfer semantics to the machine. It is enhanced with a fourth color (violet), which is used to represent missing values.
- TLP based on the thermometer: This is a new method to automatically determine the color of each cell using the knowledge and semantics given by the thermometer.
- Data-driven 2nd generation indicators (DD2gI): This is a new methodology that enriches the data-driven 3rd generation variable creation presented in [1] with the introduction of the thermometer, combined with clustering and the traffic light panel.

- Index of potential explainability: This is a new index based on the Lebart test values for qualitative variables computed versus location. It is used as the metric to select candidate variables inside TFSM.

To summarize, a new feature selection methodology is created based on multiview clustering applied with the thermometer tool.

The proposal is applied to a real case study from the INSESS-COVID19 database [2], a database resulting from research on the impact of the first COVID-19 lockdown on vulnerable Catalan populations. That study started collecting data from vulnerable citizens from Catalonia between April and 6 December 2020, and by 15 December 2020, the INSESS-COVID19 general report was publicly presented in front of governmental authorities [3].

Our proposal is applied to discover what the specific results were for local territories, and it is an improvement to the methodology proposed in [2]. Concretely, the second- and third-generation variables announced in [3] are now built on the basis of a new tool proposed in this work called the thermometer, which allows us to extract more information from the database. The INSESS-COVID19 database contains variables from several topics; thus, multiview clustering is convenient. In order to mitigate the bias provided by unbalanced views (with different numbers of variables), a new approach is developed, combining some feature selection tasks with the creation of new variables. The new methodology introduces a new tool called the thermometer, which is formalized here for the first time and allows the injection of semantics into the automatic interpretation of clusters, which are based on traffic light panels. This paper also shows how using the proposed methodology based on AI tools and methods is useful in supporting decision-making.

The structure of this paper is as follows. After this introduction in the Section 1, in Section 2, the materials and methods are shown. Within this section, Section 2.1 is dedicated to the state of the art; research related to feature selection and multiview clustering is presented. Sections 2.2–2.5 present tools that might be known to understand this paper, and Section 2.6 contains the main contributions of this paper, the thermomenter and the creation of TLP, which is based on the thermometer. Sections 2.6.1 and 2.6.2 describe the creation of the data-driven second-generation indicator (DD2gI). Section 2.6.3 describes the territorial feature selection method, and Section 2.6.4 describes its validation. In Section 3, our methodology is applied to a case study (the INSESS-COVID19 project), giving examples and details of the methodological development, such as the creation of the data-driven second-generation variable in the coexistence unit block. The practical validation is also shown (see Section 3.2.4). Section 4 draws conclusions.

In this paper, we present a new methodology that can be used in decision making and which has real-world applications. The data used in this paper are from the INSSES-COVID19 project. These reasons have moved the authors to present this paper to the journal.

2. Materials and Methods

2.1. State of the Art

2.1.1. Multiview Clustering

Multiview clustering deals with high-dimensionality data, as explained in [4], where nutritional patterns are identified through multiview clustering, or [5], where a new methodology is presented to discover and understand complex patterns with multiview clustering techniques. In this approach, the variables are split into several groups according to different themes or topics referred to in the dataset (such as anagraphics, working situations, biomarkers, opinions, etc.). Each group or view is analyzed independently from the other, although some views can be grouped into a bigger group and analyzed together; this helps when the topics of several views are related or when views contain few variables, so several blocks that make sense are grouped. Then, subjects are clustered under each group of variables or views.

Once all views are clustered and analyzed, a new qualitative variable appears. In multiview clustering, to make a general cluster, the class variables resulting from each view

are used. Multiview clustering techniques are used in [6], where variables are split into two independent groups according to their meaning.

2.1.2. Feature Selection

Which are the best variables to use in the final clustering? This is the main question to solve before clustering variables. This problem has been researched several times, and there are several papers summarizing proposed methods.

Ref. [7] provides a comprehensive and structured overview of recent advances in feature selection research. Additionally, Ref. [7] revisits feature selection research from a data perspective and reviews representative feature selection algorithms for conventional data, structured data, heterogeneous data, and streaming data. For each type of data, several methods are presented. Similarity-based methods, information-theoretical-based methods, sparse-learning-based methods, and statistical-based methods are defined for conventional data. For structured features, there are some methods for group feature structures, tree feature structures, and graph feature structures. Methods for feature selection algorithms with linked data, multi-source feature selection, and feature selection algorithms with multiview data exist in the case of heterogeneous data. Algorithms with feature streams and algorithms with data structures with data streams exist for streaming data.

Out of all the methods described in [7], the methods with more impact are the following:

- The Laplacian score, which was used in [8], where the methodology is based on the observation that data points from the same class are close. It was compared to one supervised and another unsupervised procedure. ReliefF, which was used in [9], is a similarity-based method for conventional data. Ref. [9] investigated and discussed how and why ReliefF methods work and all their characteristics (proprieties, parameters, dependencies, scalability, robustness, etc.)
- Mutual information feature selection, which was used in [10], which investigated the application of this kind of feature selection method to evaluate a set of variables and select an informative subset to be used for a neural network classifier. The minimum redundancy maximum relevance method was used in [11], which studied how to select good features as a function of the maximization of the maximal statistical dependency criterion based on mutual information. Conditional mutual information maximization was used in [12], where features were selected if they maximized their mutual information with the class to predict the conditions of any feature already selected. This method ensures that selected variables are individually informative. A fast correlation-based filter was used in [13], where the authors proposed a filter method that identified relevant features as well as redundancy among relevant features without a pairwise correlation analysis. All these methods are information-theoretical-based methods for conventional data.
- Feature selection with a lp-norm regularizer was used in [14], where the authors proposed a new method for making estimations in linear models by minimizing the residual sum of squares. This method is a sparse-learning-based method for conventional data.
- In Ref. [15], which presents statistical techniques applied in geology, the T-score appears as a statistical-based feature selection method for conventional data.
- Group lasso was used in [16], where efficient methodologies were proposed for the extensions of lassos for variable selection and were shown to improve the performance. The overlapping sparse group lasso method used in [17] proposed a new penalty method that leads to sparse estimators when it is used as a form of regularization to minimize the empirical risk. Both are methods of creating group feature structures for structured features.

In the field, unsupervised methods, such as those in [18], have tested how powerful feature-weighting techniques are for predictive tasks within different domains.

2.1.3. Selecting Discriminant Variables

It is necessary to take variables with significant differences between modalities for good feature selection. In sustainability datasets, locations are usually a modality of one variable in the dataset. To know if a variable behaves in a different way in one territory compared to another, it is necessary to use the test value proposed in [19]. This new methodology is based on [20], the original book where Lebart presented his test, which was mostly used during the profiling part when the obtained classes from a clustering should be discovered.

2.2. Creation of New Variables

Preprocessing is an important step in the data mining process because, by using accurate methods, greater quantities and quality of information can be obtained. In [21], preprocessing techniques, such as transformations and the creation of new variables, are presented. One preprocessing technique involves creating new variables corresponding to the reasoning behind parameters from the expert. Several methods of creating new variables are presented in [1]. Indicators and counts are examples of these methods. Additionally, in [22], a simplified index for water quality is presented that uses a new variable as a mathematic operation of other variables. Additionally [23], binary variables are created in the medical field. Ref. [24] presents a method of transforming correlated nonnormal variables to independent standard normal ones.

Many databases are full of variables corresponding to the same topic, which are called a block in this paper. Ref. [1] presents a method of creating a new data-driven third-generation indicator.

2.3. Clustering

Clustering is applied when which group the data belongs to is unknown and it is necessary to find those groups. As said in [25], clustering is a process of grouping data into classes or clusters so that data from the same cluster are quite similar and data from different clusters are different. Nevertheless, when clustering is applied, several elements should be decided. For example, partitioned or hierarchical methods could be applied. The authors of one paper applied a hierarchical ascendant method with Ward's aggregation criteria [26]. Ward's aggregation criteria are used in several studies, such as in [27], which presented unsupervised learning methodologies related to the field of the simulation of data, or in [28], where Ward's method was used for an application related to COVID-19. Moreover, related to hierarchical clustering, in [29], several algorithms for hierarchical clustering are presented.

In order to be able to perform this grouping of data, it is necessary to consider the measurement of distance in front of two objects. There are a large number of distance measurements. In this project, the Gibert mixed metric is used [30] since the INSESS-COVID19 instrument combines numerical and qualitative data together.

Additionally, clustering can be conditional or not. In this paper, conditional clustering is used in our proposal of our new methodology. It is used when individuals with a certain characteristic should be in the same clustering. In projects where the main object is making territorial groups, conditional clustering is a good option. Ref. [31] is a paper that presents several examples of conditional clustering, the new methodology to be applied.

2.4. TLP (Traffic Light Panel)

The TLP technique is a profiling method introduced in [32] to assist experts in the final process of conceptualizing the obtained classes. The steps to build a TLP from a clustered database are as follows:

1. Represent all variables versus the discovered variables in a class panel graph (CPG), a compact graphic tool with the conditional distributions of the variables vs. classes where the particularities of each class are easily shown.

2. Calculate the mean, standard deviation, minimum, and maximum for numerical variables and the absolute and relative frequencies for qualitative variables.
3. Using the results from Steps 1 and 2, identify variables or combinations of variables with specific ranges of values that distinguish the class from others.
4. Assign qualitative levels (high values, medium values, and low values) to the variables identified in step 3 by detecting the area where the mass of the distribution is placed.
5. Perform a TLP for the variables using the qualitative values assigned in Step 4.
6. Show the TLP to an expert and ask him to select a label for the class, which could be good or positive, neutral or medium, and bad or negative. The expert is conceptualizing the class up to this step on the basis of the traffic light panel. Colors are assigned in accordance with the interpretive codes of the expert. Green is assigned to positive or good values of the variable, yellow is assigned to medium or neutral values, and red is assigned to negative or bad values. The context and meaning of each color must be related to some latent concept of the domain that allows the association between the variable polarity and the idea of improvement or worsening. The authors of [33] propose two basic ways to assign colors to the scale of the variable.
    a. Direct color coding (red–yellow–green) for low–medium–high values.
    b. Inverse color coding (green–yellow–red) for low–medium-high values.
7. Perform significance tests (ANOVA, Kruskall–Wallis, or χ^2 independence tests) to assess the relevance of differences for the variables implied in the above steps.

A good method for TLP validation is to see at a glance how each class is distinguishable from another class. Each class must be different from another class, and two classes cannot have the same color for each variable.

2.5. Annotated TLP (aTLP)

In the TLP described in Section 2.4, only 3 colors are shown. Red always represents negative or bad values, yellow represents neutral or medium values, and green represents positive or good values. Therefore, in the original TLP, the associated uncertainty (measured by the variation coefficient) associated is omitted, and related ambiguities appear. Ref. [34] presented annotated TLP (aTLP) to improve the interpretation of TLP. Building aTLP requires the quantification of the homogeneity of the local distributions inside the classes because the tone of the color changes as a function of the heterogeneity. Therefore, a darker cell represents higher heterogeneity among the class individuals in that variable, and more uncertainty about the standard behavior of that variable in that class is propagated. Therefore, the darker an aTLP is, the higher the uncertainty related to the final profiles is; also, purer colors appearing in a TLP indicates that less variability in the central tendency of the variable is present. A color-based model for the automatic calculation of the saturation of each color is presented in [27].

2.6. Methodological Contributions

2.6.1. Thermometer

In ref. [35], two methodologies to automatically build the TLP are presented. These methods are based on different central trend statistics of variables inside the class. In this section, a new method based on expert knowledge is shown.

A thermometer (T) is a knowledge acquisition tool that allows the representation of the semantics associated with a variable in a formal way so that it can be injected into further data analysis methods. A T represents a symbolic abstraction of these semantics according to two main principles:

(a) There is a latent reference concept that can guide the evaluation of the variable values as promoting or not promoting the individuals regarding these latent concepts (i.e., water quality, the goodness of a care system, the availability of services…). This latent reference concept is aligned with the goals of the analysis.

(b) A set of traffic light colors (red (r), yellow (y), and green (g)) is associated with the semantics of the values of the variables according to the latent reference concept. For example, variables indicating dirty water will be associated with red for water quality problems, and clean water will be associated with green. Violet will be used for missing values.

The formal representation of a thermometer is described in the following:

- $\mathcal{I}$ is a set of individuals $\mathcal{I} = \{i_1, \ldots, i_n\}$ described by K variables $\{X_1, X_2, \ldots, X_j, \ldots X_K\}$.
- $D_k = \{m_1, m_2, m_3, \ldots, m_k\}$ is the set of modalities for a qualitative variable X_k.
- $T = \{t_1, t_2, t_3, ..., t_K\}$ is the available thermometer panel, where t_k, $k \in \{1{:}K\}$ is the thermometer for the variable $X_k \in K$.
- When X_k is qualitative, $t_k = \{(m_1;q_1),(m_2,q_2), \ldots ,(m_k,q_k)\}$, where:
  a. $M \in D_k$ is a modality of variable X_k;
  b. q_k is the color assigned to m_k.
- When X_k is quantitative: $t_k = \{r_1, r_2, o\}$, where:
  a. r_1 is a numerical value for X_k such that $\min(X_k) \leq r_1 \leq \max(X_k)$;
  b. r_2 is a numerical value for X_k, such that $r_1 \leq r_2 \leq \max(X_k)$;
  c. o is the semantic polarity of the variable ($o \in \{\text{direct, inverse}\}$). It represents a direct association of the variable's meanings with the traffic light colors or their inverse meanings (high values of numerical variables can link to red if they measure water pollutants or can link to green if they measure, for example, biodiversity in water quality problems).

A. NUMERICAL VARIABLES

The thermometer of a numerical variable indicates the cutpoints of a cutoff, thus building three intervals in a variable's range, which can be associated with three zones of color according to the previous experts' knowledge and thermometer principles.

By convention, due to the strong relationship that this visual model maintains with traffic lights, only a maximum of three zones with the colors red, yellow and green are allowed, and violet is used for missing values. This restricted model with three colors is oriented to take advantage of all implicit interpretation codes associated with traffic lights so that the concept's induction from the analysis results is empowered. The colored areas are arranged in such a way that:

- The first color zone ranges from $\min(X_k)$ to r_1;
- The second color zone ranges from r_1 to r_2;
- The third color zone ranges from r_2 to $\max(X_k)$.

The idea is that the expert can indicate on T three areas where the semantics of the variable move between low, normal, and high values, and the cutpoints r_1 and r_2 determine the values where the variable changes meaning.

From this first structure, a second semantic layer of the variable is transferred to the system in a second component: the color, which associates green with the most benevolent strip in terms of the interpretation of the variable and red with the least benevolent, always according to the expert on the topic studied.

The association between colors (green, yellow, and red) and qualitative levels (low, medium, and high) can be done in two ways:

- Direct: low values are red, and high values are green.
- Inverse: green values are low, and high values are red.

Determining the direct or inverse color association depends on the semantics of the variable and their values and is entirely the choice of the expert. Keeping the expert in the loop becomes critical for the construction of the thermometer.

Figure 1 visualizes a t_k from a numerical variable. It shows the following elements:

- Variable name: Name of the quantitative variable represented in the thermometer.
- Minimum: Minimum value of X_k observed in the sample.

- Maximum: Maximum value of X_k observed in the sample.
- Scale: A graduated axis with possible values of X_k.
- r_1: The upper bound of the first color zone.
- r_2: The lower bound of the third color zone

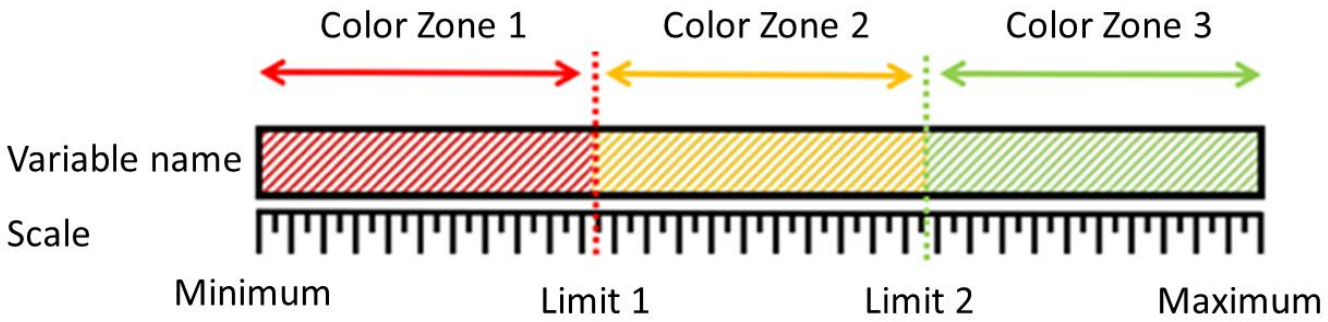

Figure 1. Numerical variable thermometer design.

When o is direct, the green value is in the third zone; if it is inverse, it is in the first zone.

The result obtained using this technique is a very powerful visual model that provides a lot of information to the system and the profane agents in the damaged nature of the data through a very visual, intuitive and intelligible representation.

Due to its compact shape, in a small space, thermometers of several variables can be shown together as rows of a single panel that, at a very quick glance, allows the expert to understand the conceptualization of the variables in a very intuitive way. The thermometer is a technical tool that collects the semantics of the variable and facilitates the creation of a new qualitative variable, catching the semantics associated with the low, middle or high values of a variable through symbolic colors red ®, yellow (y), green (g), and violet (v). It can be very often used as an auxiliary tool to automatically produce an interpretation of data analysis results.

B. QUALITATIVE VARIABLES

When the variable is qualitative, the thermometer provides a direct re-coding between the original modalities of the variable and the symbolic colors.

As there are no cutoff zones, thermometers for qualitative variables are designed in such a way that all the modalities can take any of the 3 colors. In this way, it is guaranteed that the expert can assign semantics to each modality separately due to being allowed to assign the same color to several modalities. For ordinal qualitative variables, the modalities should be represented with the correct ordering, and colors should be assigned following this order (see Figure 2).

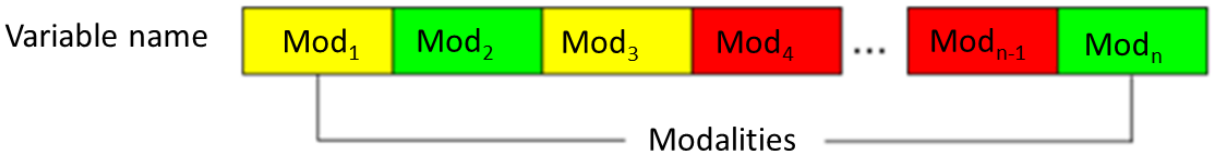

Figure 2. Qualitative variable thermometer design.

2.6.2. Creation of TLP based on Thermometer

In Section 2.4, the process of creating a classic TLP is shown. Here, the original methodology is updated to incorporate the semantic information provided in the thermometer. Let $\{X_1, X_2, X_3, \ldots, X_K\}$ be the qualitative or quantitative set of variables to be represented in a TLP, let T be the available thermometer for the same set of variables, and let P be the class variable, which is the target categorical variable to be explained in the TLP. The generation of the ***TLP based on T*** is composed of the following vector:

1. The discretization/re-coding phase: Create Z_k, a new qualitative variable resulting from the re-coding or discretization of X_K according to its original type and the thermometer information:
   a. If X_k is quantitative, create $Z_k = dis(X_k, t_k)$ by discretizing X_k according to the cutoff values indicated in the thermometer and the associated colors.
   b. If X_k is qualitative, create $Z_k = rec(X_k, t_k)$ by re-coding X_k according to the colors given in the thermometer for each modality.
   c. If $t_k \notin T$, then assign yellow to all values of X_k; the user can manually edit this.

   Discretization (X_k: quantitative):
   For all $i \in [1 : n]$, where x_i is the value of X_k for individual I, the value z_i of Z_k is as follows:

- If x_i is a valid value of X_k,
  a. If $x_i \leq r_{1:}$
     i. If o = direct, then z_i = "r";
     ii. If o = inverse, then z_i ="g";
  b. If $(x_i > r_1) \wedge (x_i \leq r_2) \rightarrow z_i$ = "y";
  c. If $x_i > r_2$:
     i. If o = direct, then z_i = "g";
     ii. If o = inverse, then z_i = "r".
- If x_i is missing, then z_n = "v".

   Re-codification (X_k: qualitative):
   For all $i \in [1 : n]$, where x_i is the value of X_k for individual I, the value z_i of Z_k is as follows:
   Given $t_k = \{m_m, q_m\}_{m=1:nk}$,
   a. If ($_{xi} = m_m$), then $z_i = q_m$;
   b. If x_i is missing, then z_i = "v".

2. The cross-matrix creation phase can be represented as follows:

$$M_k = P\ x\ Z = \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ \vdots & & & \\ n_{c1} & \cdots & n_{cq} & \cdots \end{bmatrix}, \quad (1)$$

where D_Z = {r,g,y,v} are the colors associated with the M_K columns, and given c $\in P$ and q $\in D_Z$, the element n_{cq} is the number of individuals of class c with Z_k = q.

$$n_c = \sum_{q=1}^{4} n_{cq}, N = \sum_{\forall c \in P} n_c \quad (2)$$

3. The color assignment phase. Let Fc $\in$ Mk be a row of the above matrix. The cell color is denoted as SC and is expressed as follows:
   a. Binary qualitative variables:
      i. If argmax(F_c) = 4, then S_C = v.
      ii. If argmax(F_c) = 3, then S_C = y.
      iii. If argmax(F_c) = 1:
         1. If $n_{c1}/n_c \geq \gamma$, then S_c = r;
         2. If $n_{c1}/n_c < \gamma$, then S_c = y.
      iv. If argmax(Fc) = 2:
         1. If $n_{c2}/n_c \geq \gamma$, then S_c = g;
         2. If $n_{c2}/n_c < \gamma$, then S_c = y.

   where $\gamma \in [0,1]$ and determines the threshold proportion of a modality to be considered a non-yellow color. This is required because the binary variables have only

two modalities and represent a basic dichotomy, and in a given class, the number of red or green elements has to be transformed into a single-colored cell. The default value for γ should be 0.5 so that more than 50% of the elements in a class assigned the green color would comprise a green cell in the TLP. The parameter γ allows more flexibility to deal with false positives and false negatives and provides the possibility of continuing to assign the yellow color to a cell up to a higher proportion of green (i.e., for $\gamma = 0.7$, the algorithm will not assign the green color if the class is composed of less than 70% green elements).

b. Other variables:

i. If card[$argmax(F_c)$] = 1, then $S_c = q_{argmax(F_c)}$.
ii. If card[$argmax(F_c)$] > 1 $\wedge$ ($argmax(F_c)$ = 3), then S_C = y.
iii. If $n_{c^*}/n_c > \gamma$, then S_C = v.

where $q_{argmax(F_c)}$ is the assigned color according the position of q in D_z = {r,g,y,v}.

2.6.3. Creation of Data-Driven Second-Generation Indicators (DD2gI)

In [1], several methods of building new indicators with the original data are presented. Those methods used mathematical formulas to build those indicators. In [1], the methodology was based on using specific domain knowledge provided by the experts to build new variables as a combination of the original variables. In this paper, an additional methodology is introduced to build a *data-driven second-generation variable*, whereas the original methodology presented in previous works will be known from now on as *the knowledge-based second-generation variable*.

In several databases, temporal variables are present and usually are represented as several columns of the dataset, each with a replica of the target variable at different timestamps. In other occasions, some variables are strongly related to each other because they respond to the same concept (such as a set of variables for different types of social subsidies represented in a set of binary variables). If these variables should be inputs of a clustering process, the risk of biasing the results is high, as the same "concept" is represented with more dimensions in the dataset, and thus, the weight of that concept in the cluster formation would increase. To avoid this kind of bias, a *data-driven second-generation variable* is created according to the following steps:

1. **Select the component variables to be synthesized** in a single data-driven variable: Experts should identify the subsets of variables in this situation and determine the variables that are to be components of the new variable.
2. **Cluster the selected component variables**: Cluster the selected variables with Ward's aggregation criteria [26], the chained reciprocal neighbor algorithm presented in [36], and Gibert mixed metrics. Let P be the new class variable obtained. The resulting classes are identified by the software with a number. In this form, the class variable cannot be interpreted by itself; a postprocessing technique is used to induce labels for each cluster and create the new qualitative variable with meaningful modalities.

This is addressed in the following steps.

3. **Create a CPG**: Build a CPG from all components chosen in Step 1 versus = the $\mathcal{P}$ identified in Section 2.
4. **Create a thermometer** for the component variables selected in Step 1: This should be accomplished together with the expert in the field, according to Section 2.6.1.
5. **Create a TLP based on the thermometer**: This should be accomplished using the methodology described in Section 2.6.2.
6. **Create the interpreted new indicator** $\mathcal{P}$: This should be accomplished by labelling modalities of $\mathcal{P}$ according to the information provided by the TLP/aTLP and the joint representation of TLP/aTLP over the CPG. These tools show the differences and commonalities of the different variables in each class, so the domain experts can induce proper labels for all the classes in P according to their main characteristics and

can summarize the main concept behind each class. There is a bijective relationship between P and $\mathcal{P}$. Thanks to the TLP and the thermometer, the class variable becomes an interpretable new qualitative variable with modalities having semantic meaning.

7. **Name $\mathcal{P}$:** Associate a label to the variable (often the name of the concept) and provide a description of the variable itself and each of its modalities to fix interpretation.
8. **Adding $\mathcal{P}$ to the general database**: This step enlarges the set of variables with this new variable. $\mathcal{P}$ becomes a new column of the dataset, indicating a labeled cluster for each individual with the structure of an ordinary qualitative variable.

Figure 3 shows a graphical summary of this process.

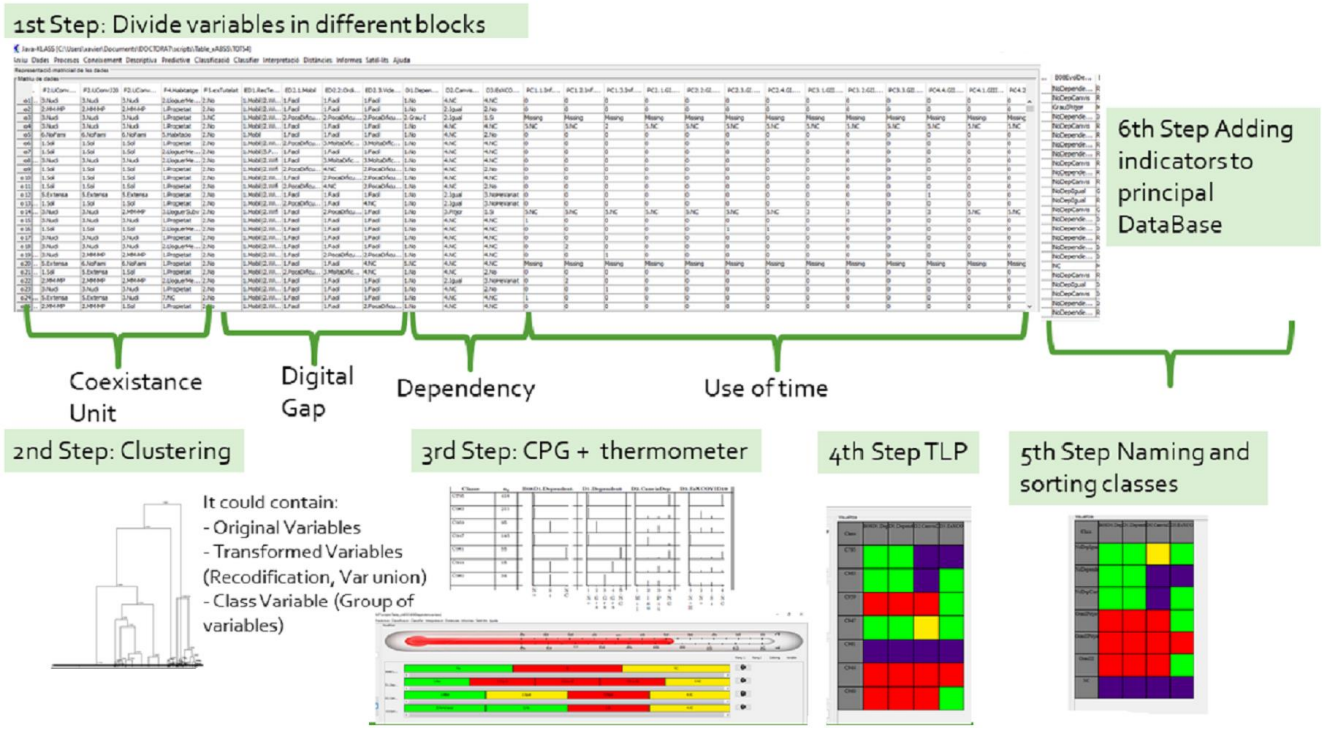

Figure 3. New data-driven 3rd generation new indicators outline.

2.6.4. Territorial Feature Selection Method (TFSM)

After finding the clusters of a view and interpreting them, feature selection is needed to make coherent territorial groups. The focus is on finding which variable is the best to represent the view in the final cluster, so one variable per block will be allowed in order to avoid the overrepresentation or misrepresentation of one topic.

At this point, each block has its own indicator with a balanced number of component variables. Each indicator is composed of original variables and second-generation variables. This is the appropriate time to select the variables to make a general cluster, which should have one variable per block. The methodological contribution to select the appropriate variable of each block is based on the Lebart test values presented in Section 2.1.2. The main goal of this method is to obtain the selected variable per block to be used in the final global cluster where $\mathcal{I}$ is a set of individuals $\mathcal{I} = \{i_1 \dots i_n\}$ described by K variables $\{X_1, X_2, \dots, X_j, \dots X_K\}$.

The steps are as follows:

1. **Select the location variable:** In this case, the territorial variable should be selected from the original database. This is an informative variable in the original dataset that does not belong to any block. The territorial variable X_{Loc} is a qualitative variable where the modalities are locations (L) $D_{Loc} = \{l_1, l_2, \dots, l, \dots l_L\}$, *Loc* ϵ 1:K.
2. **Select the block:** In Step 1 of the creation of new data-driven third-generation indicators, some blocks were created. Now, one block should be selected.

3. **Select the candidate variables:** At this moment, the block might contain original variables, second-generation variables, and the data-driven third-generation indicator. Only the third-generation variable and its component variables are selected as candidate variables. The candidate variables are $\chi = \{X_k \text{ tq } k \in 1{:}K\}$ with modalities $D_k = \{m_1, m_2, \ldots, m, \ldots m_{n_k}\}$, $m \in D_k$
4. **Compute a ranking of variables according to their capacity to explain the territorial distribution:** Evaluate the candidate variables to determine their selection. Repeat the following steps for each candidate variable pre-selected in Step 3.
   a. Calculate Lebart test values: Using the methodology from Section 2.1.2, all Lebart test values are calculated for each qualitative candidate variable against *Loc*. For each X_k qualitative, the output of this process is a table V_k with $l \in D_{Loc}$ in rows and n_k columns, named $M_{m\ (m=1:\ n_k)}$, where v_{lm} are the p-values of the Lebart test value of M_m *versus* $l \in D_{Loc}$.
   b. Create a new variable S_k with $l \in D_{Loc}$ in rows.
   c. Calculate the ratio of significant locations for the variable X_k using the information provided by S_k,

$$R_k = \frac{card\{s \in S_k : s > 0\}}{L} \tag{3}$$

   This indicator assesses the percentage of locations that can be characterized by some modalities of X_k. We will be interested in the variable X_k that provides significant modalities to as many locations as possible, meaning that the same variable can explain a bigger part of the territory.
   d. Given X_k, and its corresponding V_k, create Table Π_k with $l \in D_{Loc}$ in rows and n_k columns, named $M_{m\ (m=1:\ n_k)}$, where π_{lm} corresponds to the empirical probability of X_k*conditionedtoagivenlocation* for significant cells:

$$\pi_{lm} = \begin{cases} 0 & v_{lm} > 0.05 \\ \frac{card_{i \in \mathcal{I}}\{x_{ik}=M_m \text{ and } x_{iloc}=l\}}{card_{i \in \mathcal{I}}\{x_{iloc}=l\}} & v_{lm} \leq 0.05 \end{cases} \tag{4}$$

   This helps to understand where the significant modalities of a certain location cover a big portion of the location population or, on the contrary, target a limited group of individuals. In fact, having a significant modality that covers a minority is not sufficient to explain a cluster (or a location).
   e. Create a new variable Π_{Loc*k} with $l \in D_{Loc}$ in rows and

$$\Pi_{l*,k} = \sum_{m=1}^{n_k} \pi_{lm} \tag{5}$$

   $\Pi_{Loc*,k} = (\Pi_{1*,k}, ..., \Pi_{L*,k})$ indicates the proportion of individuals involved in locations with significant modalities of X_k according to the Lebart test value.
   f. Calculate the average of $\Pi_{l*,k}$ for all locations so that we obtain an estimate of the average contribution of the significant modalities of X_k to the location population.

$$\widetilde{\Pi}_{Lock} = \frac{\sum_{l=1}^{L} \Pi_{l*,k}}{card_{l=1:L}\{\Pi_{l*,k} > 0\}} \tag{6}$$

   This indicator is an estimate of the average coverage of significant modalities of X_k along the territory. For low values of $\widetilde{\Pi}_{Lock}$, the significant modalities of X_k correspond to a minority of the different locations, and the variable is not as informative as required.

g. Calculate the ***index of potential explainability*** of the variable X_k in the territory as

$$E_k = R_k \cdot \widetilde{\Pi}_{Loc,k} \quad (7)$$

The index E_k weights the average proportion of individuals involved in significant territories for a given variable according to the proportion of significant locations in the entire territory. We are interested in variables with the capacity to characterize as many locations as possible, with as big an area of coverage as possible. The relevance of the variable decreases if it is significant in many locations with sparse coverage or in few locations even though it involves a large number of individuals. This is a correction that is important in order to balance the impact in the analysis of big cities, which concentrate a large part of the population in a relatively small area, as is the case with Barcelona in the Catalan territory. E_k is defined to gain robustness with regard to an unbalanced distribution of the population in the territory or modalities pointing out exceptional groups of individuals with low presence. This prevents big cities and exceptional minorities from dominating the entire analysis. By computing the value of E_k for all candidate variables, a ranking can be built regarding the potential of X_k to explain the territorial distribution.

5. **Select the variables:** For each block, the variable with the maximum E_k is selected to represent the block in further analysis.
6. **Cluster the selected variables:** The set of variables selected in Step 5 has one representative variable for each block in the dataset. This set of variables is the input to a conditional clustering method using a hierarchical algorithm based on the Ward's method aggregation criteria presented in [23], Gibert mixed metrics, and X_{Loc} as a conditional variable. A new class variable $\wp$ results from this process. The resulting classes are identified by the software with a number. In this form, the class variable cannot be interpreted by itself; a postprocessing technique is used to induce labels for each cluster and make the new qualitative variable with meaningful modalities.
7. **Interpret and label the obtained classes**: Repeat Steps 3 to 8 of the *data-driven second-generation variable creation* process.
8. **Profile the classes:** Analyze the significance of input variables in the classes and the conditional distributions to identify the relevant characteristics of each class so that a short description of the characteristics of each class can be provided.
9. **Create maps to visualize classes:** This step is carried out to visualize the final classification.

2.6.5. Validation Methodology

In this section, the validation protocols for the different methodological contributions of the paper are presented for the following reasons:

- To validate the introduction of the thermometer into the generation of automatic TLPs: The thermometer method is used to improve the TLP. Thus, the validation methodology proposed is based on the comparison of two TLPs. One is built traditionally, where the color of a cell is decided by a human based on CPG analysis; the second, the thermometer-based TLP, is where the color of the cells is decided by the method proposed in Section 2.6.1. Both TLPs and their interpretation are shown to a group of experts in the field, who discuss which of the two is more believable or which of them provides a better understanding of the target domain.
- To validate the DD2gI (data-driven second-generation indicator) methodology: This is a semantically enriched methodology based on the data-driven methods presented in [1]. The validation methodology proposed is the same defined in the previous work.
- To validate the territorial feature selection method (TFSM): Here, two aspects are validated:
  a. The pertinence of the index of potential explainability that we propose: The results of using the proposed TFSM methodology are compared with the state of the art using a χ^2 test of the territorial variable (X_{Loc}) versus each single variable

X_k as a tool to select the most discriminant variable for further clustering steps. However, all X_k provide a significant χ^2 independent test *p*-value and are not helpful in reducing the number of variables to be used to represent the blocks in the clustering process, whereas the proposed index allows a ranking of variables from more discriminant to less and allows the variables to undergo a selection process.

b. The final classification obtained: Two classifications are compared. One results from the application of TFSM, namely $\wp$. Another results from clustering the set of third-generation variables created for each block, namely $\wp'$. There are two ways of comparing and validating if the classification obtained through TFSM is better; one method is based on graphical tools, and the other is based on numerical tools.

i **Numerical validation:** We start by calculating the Lebart test values. Using the methodology from Section 2.1.2, all Lebart test values are obtained for each $c \in \wp$ versus X_{Loc} and for each $c' \in \wp'$. Tables $V_\wp$ and $V_\wp$are computed similarly to the V_k tables presented in Section 2.6.3. Here, since the clustering has been conditioned to the locations, all individuals of a location are clustered into a single class. Therefore, $s_{lk} = 1 \; \forall l = 1 : L, k = \wp, \wp\prime$ and $R_\wp = R_{\wp\prime} = 1$ by construction so that $E_\wp = E_{\wp\prime} = 1$ and the proposed index of potential explainability cannot be used in conditional clustering methods. For this reason, we will just compute the global proportion of significant cells in $V_\wp$ and $V_{\wp\prime}$ to see which of the two partitions can explain a bigger part of the territory. Given a partition P with n_P classes, ($P \in \{\wp, \wp'\}$), the index is

$$S_P = \frac{card_{ln|l=1:L,c=1:n_P}\{v_{lc} \leq 0.05\}}{L * n_P} \quad (8)$$

This index accounts for the proportion of significant cells in a V_k table. The greater the value of S_P the better P distributes along the locations. We propose comparing $V_\wp$ and $V_{\wp\prime}$ and consequently $S_\wp$ and $S_{\wp\prime}$ in order to see which of the two partitions distributes into the territory in a more significant way into the territory.

ii **Graphical validation**: One map per classification should be drawn. The map should be painted as a function of the class that belongs to the location. The territorial cohesion of the classes will be considered for the evaluation.

3. Case Study and Results

3.1. INSESS-COVID19

3.1.1. The Project

The INSESS-COVID19 project (Identification of Emerging Social Needs as a Consequence of COVID-19 and Effect on the Social Services of the Territory [37]) focused on understanding and anticipating the overflow of social services expected after the overflow of the healthcare system because of the lockdown raised by the SARS-CoV-2 crisis (March 2020). The Catalan Social Services System was interested in quickly obtaining information about the possible new needs of vulnerable people. The INSESS-COVID19 project tries to provide answers and introduces the potential of coexistence clustering through a thermometer-based TLP to perform a prospective study that allows the identification of the social vulnerabilities of the Catalan population in a sufficiently understandable way such that the results can provide elements to support decision making and policy making for the 107 Basic Areas of Social Services (BASS) of Catalonia and the Social Services Department of the Catalan government.

The data collection started in April 2020 and ended on 6 December 2020, and the final results of the project were written in a general report [3]. It was published on the project web and presented to the Catalan Government on 15 December 2020, only 9 days after data

collection closed. A general report was delivered to the government and distributed to the 107 BASS (basic areas of social services) all over Catalonia, containing information from all BASS together. Nevertheless, the ideas of that report did not provide elements to support decision making and policy making for each location.

All methodological details are clearly explained in [3], which also contains the project's global results. They are described with several tables, graphs and maps. Moreover, [2] defines the new techniques applied and the previous work related to INSESS-COVID19. That information was reported to the Catalan government, which was able to make global decisions quickly in 2020 while the COVID-19 pandemic was still unfolding.

This project was a finalist of the European Social Services Awards 2021, and it has contributed to the Muncunill 2021 award from the Terrassa City Council that the Universitat Politecnica de Catalunya won.

3.1.2. INSESS-COVID19 Database

Our data source was the INSESS-COVID19 questionnaire. One of the main goals of the project was to obtain information on vulnerabilities in several areas of life; thus, the questionnaire was intended to ask about them. The target life areas included in the questionnaire came from the reference conceptual model named SSM.cat (self-suffiiciency matrix) [38], an instrument to assess social vulnerability adopted by the Catalan government to be part of the new social services system (e-Social), planned as the kernel of the digital transformation of social services targeted in the strategic plan [39] and very much aligned with the current structure of primary care social services in Catalonia.

The INSESS-COVID19 questionnaire has 18 blocks regarding different life areas, each with different numbers of items, resulting in a total of 195 variables. The complete details of the INSESS-COVID19 questionnaire and its associated database were published in [2], and Figure 1 in [2] visualizes all the questions. After the initial preprocessing step, the number of variables increased to 214. Additionally, 63 knowledge-based second-generation (KB2g) (as described in [1]), 12 data-driven second-generation criteria as described in Section 2.6.3, and 19 third-generation variables (3gV) were added (as described in [1]), resulting in a total of 295 variables. Some of the variables asked the same question in three different moments of time, as described in [2]. Table 1 displays the number of variables of each type distributed per thematic block, and in step 7 of Section 3.2.2 is shown the components of all the derived variables.

This unbalanced number of variables per block means that they cannot be treated in the same way in a global analysis, especially at the moment of feature selection.

The first blocks contain personal questions, such as age and place of residence. This last one will be important in the following sections.

The answers were obtained through special workshops developed in the INSESS-COVID19 project and described in [2]. By the end of the project, the INSESS-COVID19 database contained answers from 971 vulnerable citizens from all Catalan territory.

Table 1. Number of variables related to each topic.

Block Number	Block Name	Original Variables	Preprocessed Variables	DD2gI	KB2g	3gV
B02-B03-B04	Origin	1				1
B05	Trials	2	8			1
B06	LivingCoexistance	5		1		1

Table 1. *Cont.*

Block Number	Block Name	Original Variables	Preprocessed Variables	DD2gI	KB2g	3gV
B07	DigitalGap	4	6	1		1
B08	DependentEcolution	4		1		1
B09	UseofTimes (peopleincharge)	15			4	1
B10	UsdelTempsDINt1T2	16			9	1
B11	ConvRel	3				1
B11	Violence	19	5	4	36	1
B12	Allparticipation	12		3		1
B13	Labour	17		1		1
B13	LaborBussiness	5				1
B14	Telework	8				1
B15	Economy	18		1	14	1
B15	Gethelp	14				2
B18	Health	9				1
B18	Addiction	12				1
B18	MentalHealth	9				1

3.2. Creating New Variables

3.2.1. Creation of Location Variable

The participants were asked about their municipality of living. Nevertheless, that variable is not useful for achieving our aim. In Catalonia, there are 947 municipalities. Each town belongs to one BASS, so a new variable is created in the database called BASS. The experts gave us the list of municipalities belonging to each BASS, and we created a new variable called BASS, which indicated the BASS of residence of the participant.

3.2.2. Creation of Data-Driven Second-Generation Indicators (DD2gI)

In Section 2.6.2, a new methodology for creating data-driven second-generation indicators is presented, which summarizes the main idea of creating new variables.

Below, we present a demonstration of how second-generation variables could be built using our new methodology.

1. Select the component variables to be synthesized

In the questionnaire, there are several questions providing information about the situation of respondents in January 2020, July 2020 and January 2021, so they can be analyzed together.

In this case, participants were asked about their living coexistence. Each participant answered the question: *"Who do you live with"*. The possible options were: *"1. Alone: I live alone.; 2. MM-MP: Single-mother or single-parent family.; 3. Nucleus: Family of father and mother and own children (if there are children).; 4. Regrouped: Regrouped families (children of several couples).; 5: Extended: Extended family (fathers, mothers, children, grandparents, aunts, etc.).; 6: Non-Family: I live with people who are not family."; 7. No answer*.

This question was answered 3 times, creating 3 different variables. We analyzed these variables by making them components of a new variable.

2. Cluster the selected component variables

Using the Ward's method aggregation criteria presented in [22] with Gibert's mixed metrics, selected variables were clustered, resulting in a new class variable. Figure 4 shows the resulting dendrogram.

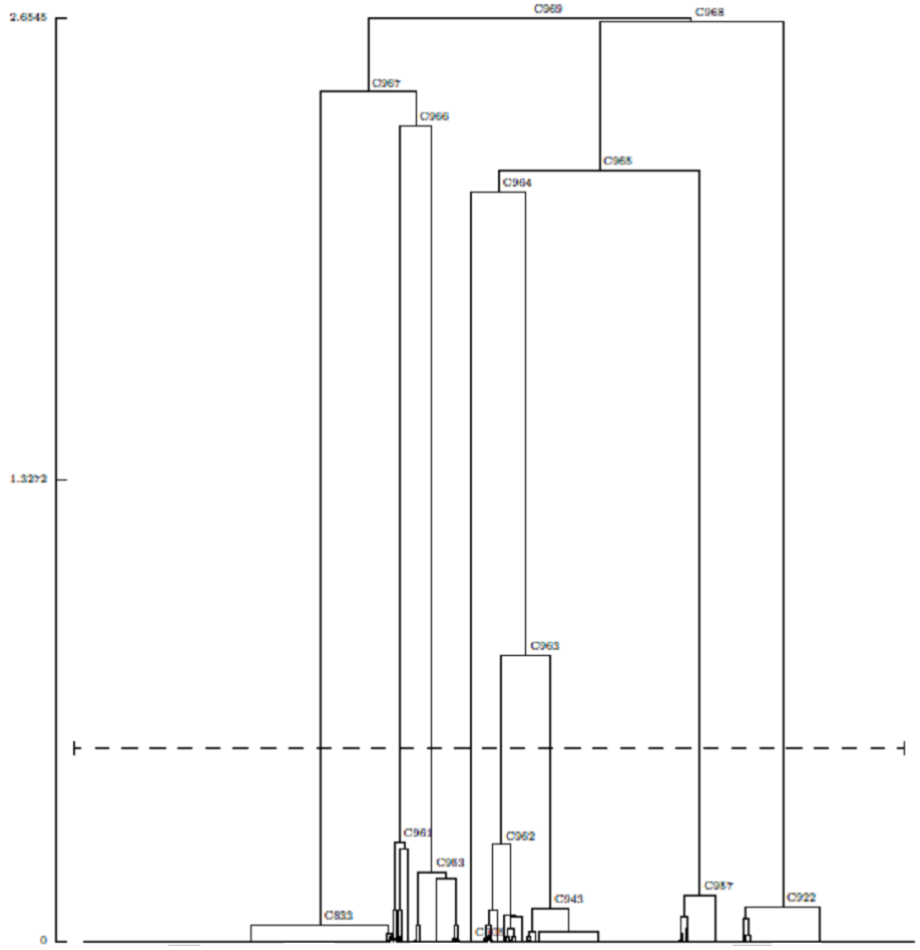

Figure 4. Coexistance Unit Dendogram.

Using Calinski–Harabasz criteria [40], the database was divided into 8 clusters.

3. Create the CPG

The CPG was created and is shown in Figure 5.

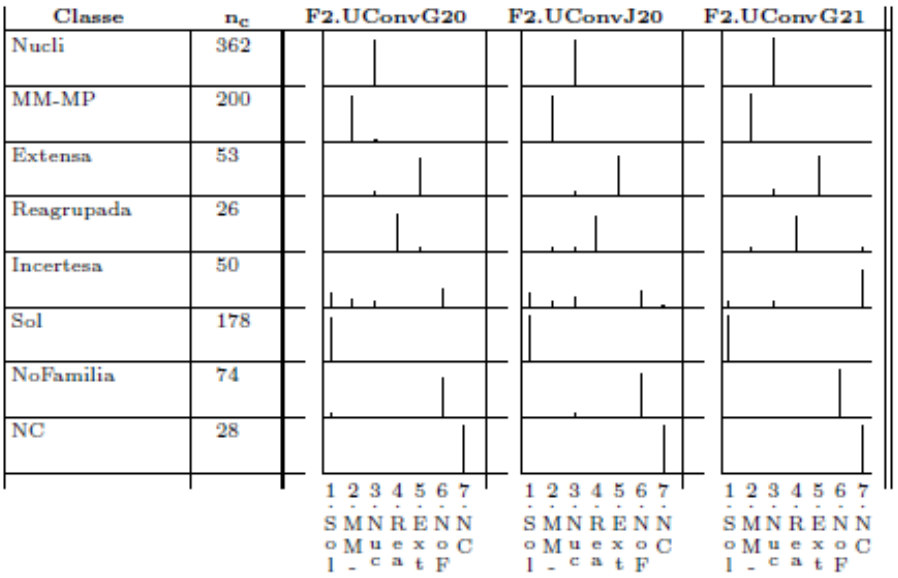

Figure 5. Living Coexistence Class Pannel Graph.

4. Create the Thermometer

Following the methodology presented in Section 2.6.1, a thermometer was created (see Figure 6), and together with the field expert, a color was assigned to each modality.

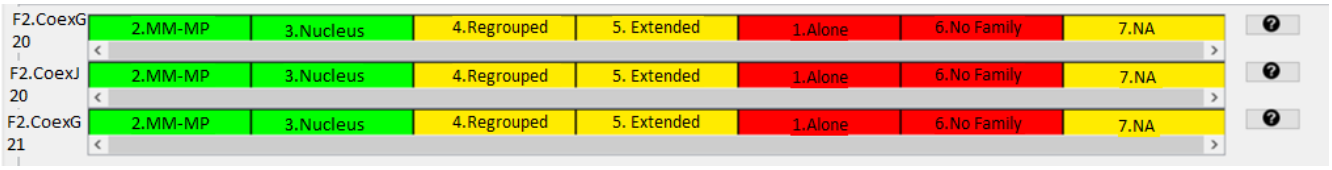

Figure 6. Living Coexistence Thermometer.

- Green was assigned to the modalities "2.MM-MP" and "3.Nucleus" because the expert considered that someone living with their nuclear family and people who love them is a positive situation. In this instance, an individual has people who can help them in case of necessity, and difficulties are easy to overcome.
- Yellow was assigned to the modalities "4.Regrouped" and "5.Extended" because the expert considered that living with people that belong to one's family group is positive. On the other hand, these kinds of families usually are composed of many members, and the apparition of conflicts between members is possible. Therefore, as this possible aspect is negative, the modality was marked yellow.
- Red was assigned to the modalities "1.Alone" and "6.NoFamily". Living with people who are not one's own family where there is, therefore, no emotional attachment is akin to living alone. It is more difficult to find help. Moreover, if one lives with people that are not one's family, it is easier for conflicts to appear.

5. Create the TLP based on the Thermometer

In Figure 7 is shown the TLP based on Thermometer

	F2.CoexG20	F2.CoexJ20	F2.CoexG21
Nucleus			
MM-MP			
Regrouped			
Extended			
Uncertainty			
Alone			
No Family			
NA			

Figure 7. Living Coexistence T-TLP.

- Nuclear: People in this group answered "3.Nucleus" for the 3 periods of time.
- MM-MP: People in this group answered "2.MM-MP" for the 3 periods of time.
- Regrouped: People in this group answered "4.Regrouped" for the 3 periods of time.
- Extended: People in this group answered "5.Extended" for the 3 periods of time.

- Coexistence: People did not know what to do in January 2021 (note that the questionnaire was answered in 2020, during the first COVID-19 wave). During 2020, they had different situations.
- Alone: People in this group answered "1.Alone" for the 3 periods of time.
- No family: People in this group answered "6.Non-Family" for the 3 periods of time.
- NA: People in this group answered "7.No Answer" for the 3 periods of time.

To summarize, there were 8 classes. Seven encompass people without changes in their living status, i.e., people that answered with the same option in January 2020, July 2020, and January 2021. In the last group, people changed their answer.

6. Create the interpreted new indicator $\mathcal{P}$ and name $\mathcal{P}$:

As this new variable comes from the variable F2.CoexMYY, where M corresponds to the month and YY to the year, it was labeled F2.CoexLab, where lab represents labeled and Uconv represents living coexistence.

7. Add $\mathcal{P}$ to the general database:

Once all steps were complete, $\mathcal{P}$ was added to the general database.

During this process, this procedure was repeated several times, as several DD2gI were created. Additionally, the new data-driven third-generation indicators presented in [1] were added to the general database. Figures 8–10 represent all variables and indicators in the database.

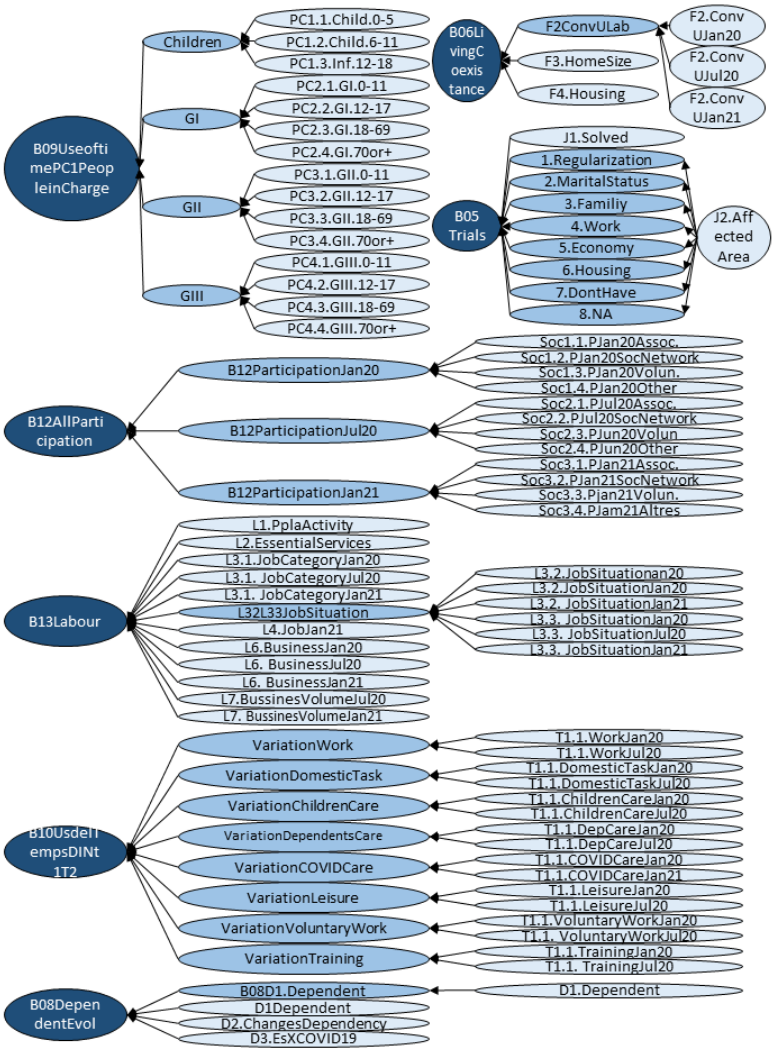

Figure 8. Indicator components I.

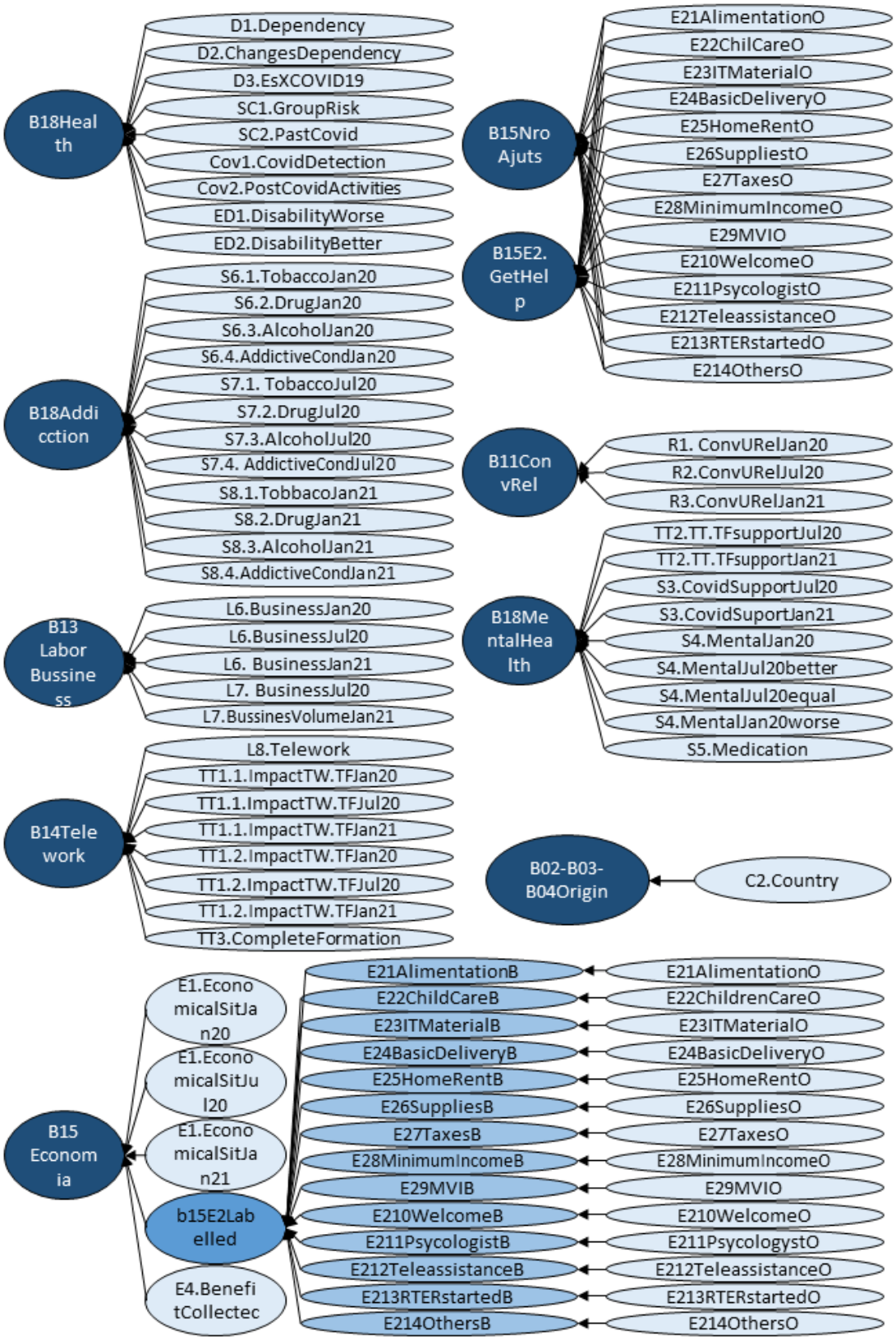

Figure 9. Indicator components II.

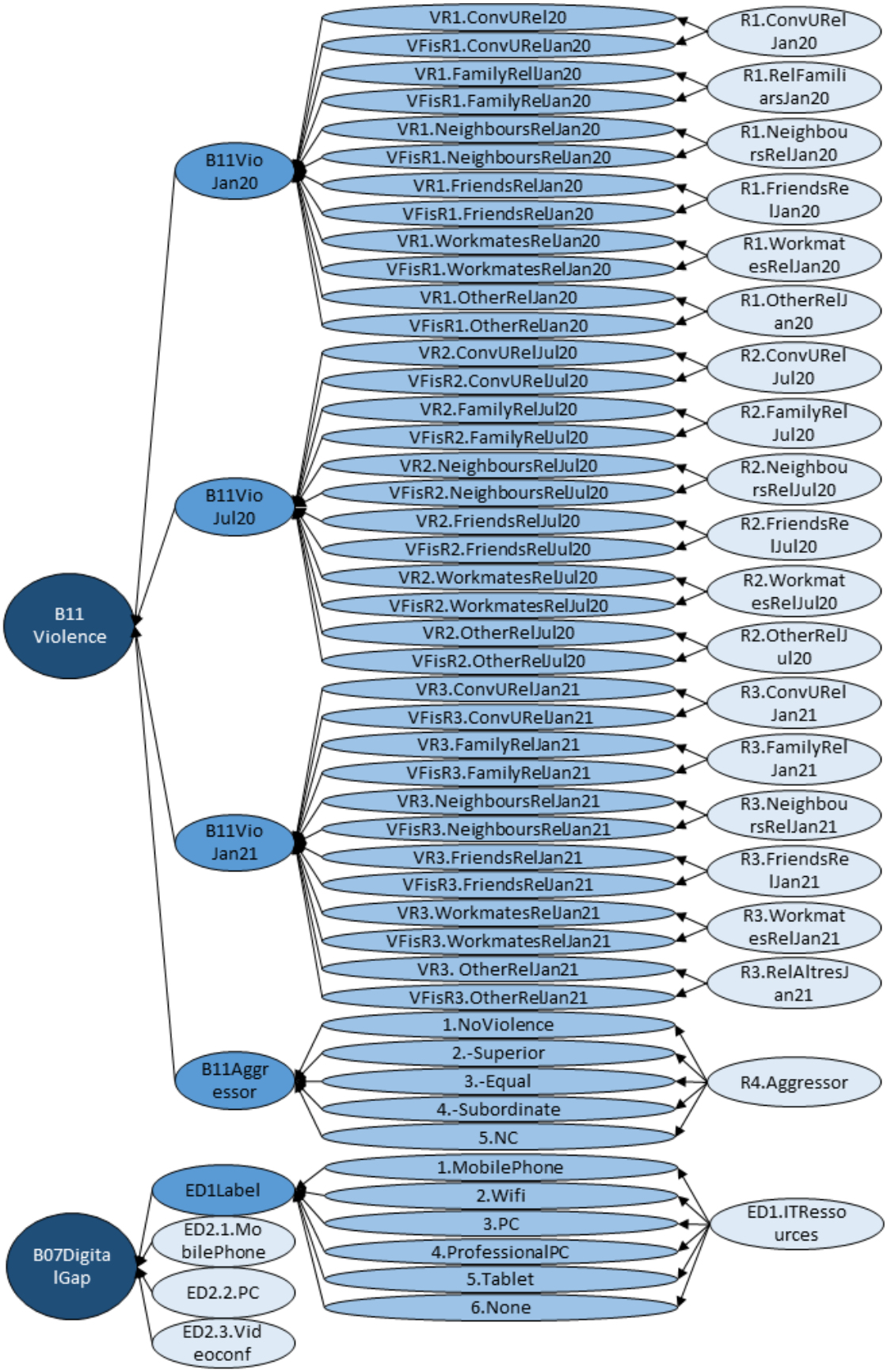

Figure 10. Indicator components III.

3.2.3. Feature Selection and Final Clustering

In the following, we show how the variables that were clustered were selected. As an example, we used Block XI: Use of Time (Caretakers) as an indicator. This indicator was composed of 4 variables. These variables indicated if the participant had children in charge or Grade I, II or III dependency people in charge. These variables were second-generation variables built using the methodology proposed in [1]

1. **Select the location variable:**

The territorial variable was the BASS. That variable was created in Section 3.2.1. This was a variable coming from the original variable for the municipality of residence.

2. **Select the block**

We used the case of Block XI: Use of Time (Caretakers), which was constructed above.

3. **Select the candidate variables**

As can be seen, 4 second-generation variables are the components of the indicator. Therefore, Children, GI, GII and GIII were the candidate variables we selected for inclusion in the final cluster.

4. **Compute a ranking of variables according to their capacity to explain the territorial distribution:**

All results were calculated using the methodology explained in Section 2.6.3. Table 2 is the result of the application of the methodology explained in step 4 of TFSM explained in Section 2.6.4.

Table 2. $R_k, \widetilde{\Pi}_{Lock}, E_k$ values for each component.

Indicator/ Component	Block XI: Use of Time (Caretakers)	Children	GI	GII	GIII
R_k	0.7	0.5	0.5	0.43	0.45
$\widetilde{\Pi}_{Lock}$	0.47	0.56	0.58	0.54	0.57
E_k	0.329	0.28	0.29	0.23	0.26

5. **Variable selection:**

In this case, the variable that represents the block was the selected variable, which, here, was Block XI: Use of Time (Caretakers).

Once all variables from the model were selected and the feature selection was carried out, the final selected variables were R3 RelConvG21(R3), B06UConvivencial (B06), S3SuportCovid19J20 (S3), B11VioG21 (B11), B02-B03-B04Origen (B02-03-04), B09UsTempsPC1FentCarrec (B09), E210AcollidaO (E210), 5Economia (5), B15E2Labelled (B15E2), L32L33SituacioLab (L32L33), B12ParticipacioTot (B12), VariacioCuraCOVID (VAR), D3EsXCOVID19 (D3), and L6NegociG20 (L6).

6. **Clustering selected variables:**

All of the selected variables and the BASS were clustered using a method of conditional clustering incorporating the Ward criteria and the mixed Gibert distance. The resulting dendrogram is shown in Figure 11.

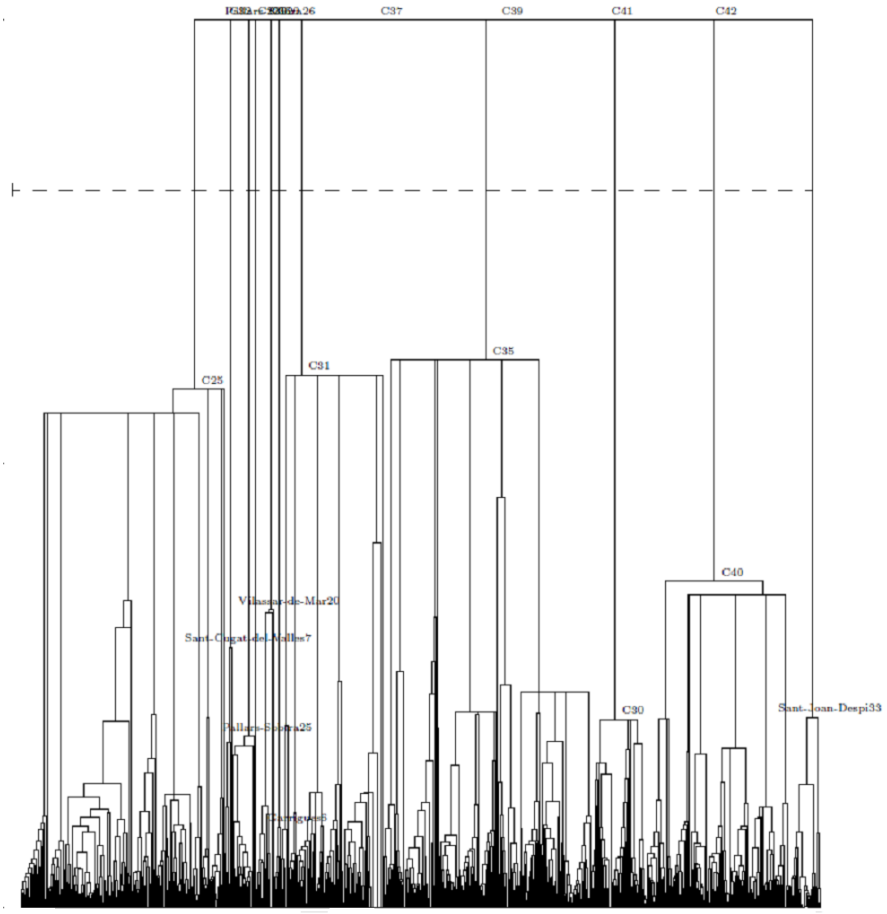

Figure 11. Final Dendogram.

Using the Calinski–Harabasz criteria [39], the database was divided into 11 clusters.

7. Creating CPG

After the dendrogram, a CPG was created as it is shown in Figure 12

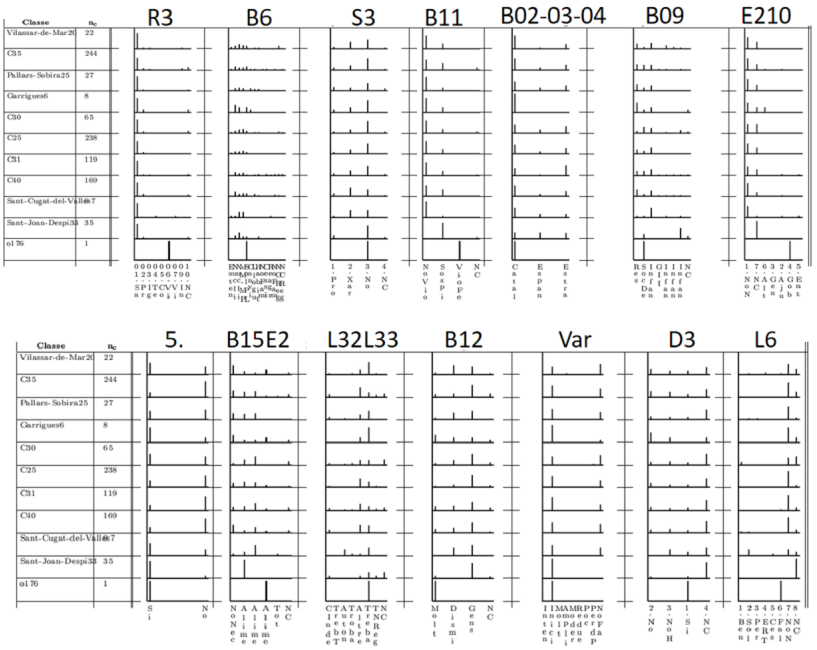

Figure 12. Final CPG.

8. **Creating Thermometer**

The final thermometer was then built. The variables that were selected were indicators or indicator components. In case the selected variable was an indicator component, the thermometer built for that item was reused in this step. In case the selected variable was an indicator, a new thermometer was built using expert knowledge and the resulting TLP to build the third-generation data-driven indicator.

Every single variable has its own thermometer. Figure 13 shows the thermometers for all the variables.

Figure 13. Final Thermometer.

9. **Create aTLP based on a Thermometer**
10. **Profile the classes**

After carrying out clustering, applying the thermometer to the TLP (see Figure 14), and using the automatic conceptualization of the TLP to the natural language descriptions described in [5], the result was 11 groups, which are described in the following section.

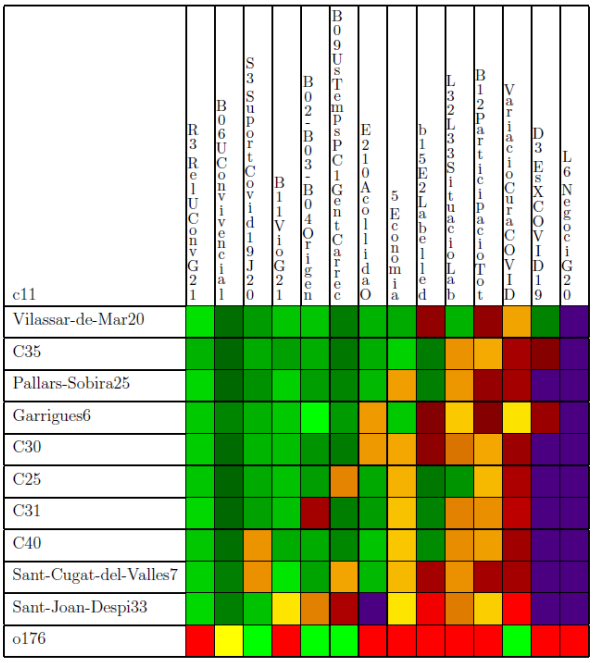

Figure 14. Final T-aTLP.

- Vilassar de Mar: This group does not present with relationship problems within their units of coexistence, and members of this group comprise families living in houses that they own or rent. Members of this group did not need psychological support due to COVID-19. They are mostly of Catalan origin. The people in charge of familities in this group do not waste any time caring for their unit. They did not need any aid due to COVID-19. Some have pending lawsuits related to the economic field. Some did not need social assistance during COVID-19. This group is comprised of working people as well as people with unconventional work situations. Their socialization with the environment has diminished or is nonexistent.
- C35: This group does not present with relationship problems regarding their unit of coexistence, and members of this group comprise familial cohabitation units (extensive, traditional, or single-parent) in houses they own or rent. They did not need psychological support due to COVID-19. They are mostly of Catalan origin. The people in charge of families in this group do not waste any time caring for their units. They did not need any aid due to COVID-19. They have no pending economic decisions, and some people need help with food. This group is comprised of people with other unconventional work situations and people who do not work. They do not participate in their environment.
- Pallars: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They did not need psychological support due to COVID-19. They are mostly of Catalan origin. The people in charge of families in this group do not waste any time caring for their family units. They did not need any aid due to COVID-19. They have no pending economic situations, and some of them need help with food. This group is comprised of people who work and who have other unconventional work situations. They do not participate in their environment.
- Garrigues: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They did not need psychological support due to COVID-19. They are mostly of Catalan origin. The people in charge of families in this group do not waste any time caring for their family units. Their situation of dependency has not diminished due to COVID-19. Some members of this group need help from social services due to COVID-19. They have pending economic decisions; furthermore, some people in this group have not needed food-related help, while others have needed food-related help. This group is comprised of people who work and people who have other unconventional work situations. Their participation in their social environment is varied. Some participate a lot, some demonstrate decreased participation, and some do not participate at all.
- C30: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They did not need psychological support due to COVID-19. They are mostly of Catalan origin. The members of this group have no people in charge of their familial units. Some of them need help from social services due to COVID-19. They have pending financial decisions, and some people have needed food-related assistance and other specific types of aid that arose during the pandemic. This group is comprised of people who work and people who have other unconventional work situations. They do not participate in their environment and have not cared for people with COVID-19.
- C25: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They did not need psychological support due to COVID-19. They are mostly of Catalan origin. Most of them have dependent children who take up part of their time. They did not need

any aid due to COVID-19. They have no pending economic decisions, and some people need help with food. This group is comprised of people who work and people who have other unconventional work situations. They do not participate in their environment and have not cared for people with COVID-19.

- C31: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They did not need psychological support due to COVID. They are of a foreign origin. The people in charge of families in this group do not waste any time caring for their family units. They did not need any aid due to COVID. They have no pending economic decisions, and some people need help with food. This group is comprised of people who work and people who have other unconventional work situations. They do not participate in their environment and have not cared for people with COVID-19.
- C40: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. Some people have needed psychological support due to experiences related to COVID-19 within their personal network. They are mostly of Catalan origin. The people in charge of families in this group do not waste any time caring for their family units. They have no pending economic decisions, and some people need help with food. Some are people with other unconventional work situations, and others are working people. They do not participate in their environment and have not cared for people with COVID-19.
- Sant Cugat del Vallès: This group does not present with relationship problems within their units of coexistence. In fact, they have good coexistence units, as they live mainly with their nuclear families or are single-parent families. Some people have needed psychological support due to experiences related to COVID-19 within their personal network. They are mostly of Catalan origin. They take care of children, who consume part of their time. They did not need to attend reception services during the COVID-19 pandemic. They have pending economic decisions, and most people need help with food. Some people worked before the pandemic but lost their jobs, while others are self-employed. Some people do not participate in their environment at all, and in other cases, their participation in their environment has decreased.
- Sant Joan Despí: This group does not present with relationship problems within their units of coexistence, and members of this group comprise familial coexistence units (extensive, traditional, or single-parent) in houses that they own or rent. They have not needed psychological support due to experiences related to COVID-19 within their personal network. They are of Spanish origin and are suspected of being victims of violence. They have children in charge who take care of them, as well as grade III dependents who take up their time. They have pending lawsuits related to economic matters, and all people need help related to food. Some people worked before the pandemic but lost their jobs. Not everyone is involved in their environment.
- o176: This group is composed of people who suffered verbal violence from their units of cohabitation but now live alone in a flat. These individuals do not need emotional support but may suffer physical violence at work and psychological violence at rest. These individuals are dependent people who live in Catalonia and have accepted the measures proposed by the Spanish government to solve the COVID-19 pandemic. They have pending lawsuits on economic issues and have received multiple grants. They worked before the pandemic and are very involved people.

11. Creating maps to visualize classes

It is easy to see from Figure 15 how groups are distributed. The distributions are as follows:

- C25: Alt Empordà, Amposta, Baix Penedès, Calafell, Mollet del Vallès, Sant Vicenç dels Horts, Solsonès, Tarragona, and Vilafranca del Penedès.

- C30: Lleida, Montcada i Reixac, Sant Andreu de la Barca, ad Sant Pere de Ribes.
- C31: Barcelona, Manresa, Masnou, el Rubí, and Tarragonès.
- C35: Alt Penedès, Bages, Baix Llobregat, CAS Garrotxa, Girona, Maresme, Osona, Pallars Jussà, Reus, Ribera d'Ebre, and Vilanova i la Geltrú.
- C40: Baix Empordà, Barberà del Vallès, Figueres, Gironès-Salt, Noguera, Pla de l'Estany, Sant Feliu de Guíxols, Selva, and Vallès Oriental.
- Garrigues: Garrigues.
- Pallars-Sobira25: Pallars Sobira.
- Sant-Cugat-del-Valles7: Sant Cugat del Vallès.
- Sant-Joan-Despi33: Sant Joan Despí.
- Vilassar-de-Mar20: Vilassar de Mar.

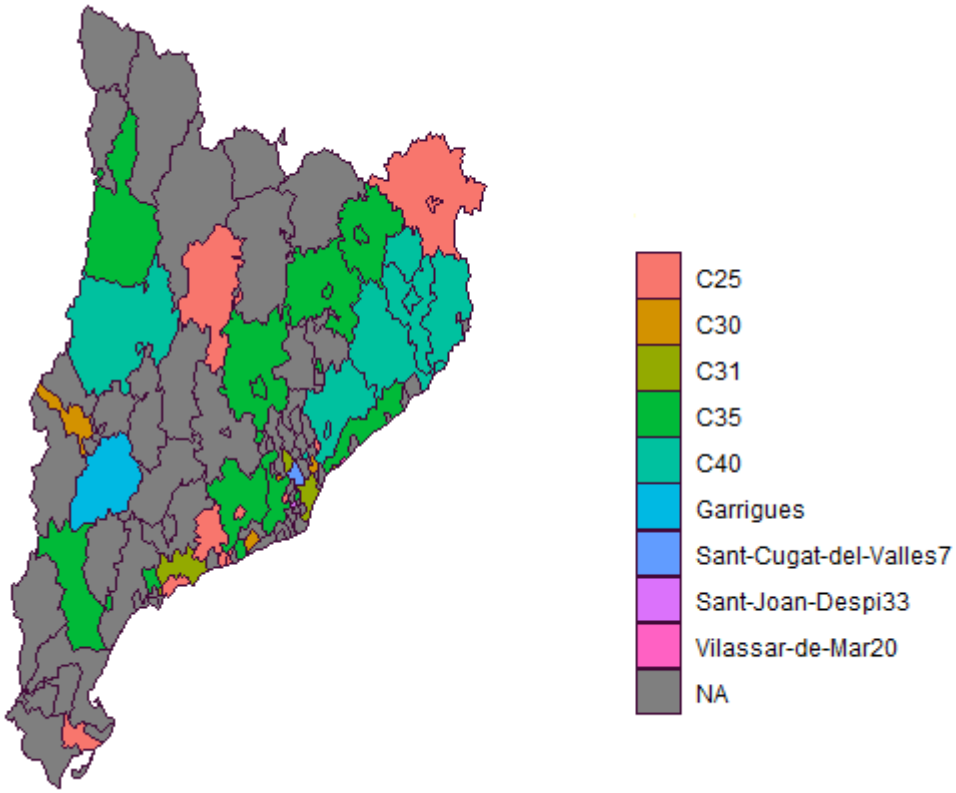

Figure 15. Territorial representation of clustering with selected variables.

Thus, we can see how the BASS were grouped.

3.2.4. Validation

Validating the DD2gI

As said in Section 2.6.4, χ^2 tests were applied for all components and indicators against BASS. Significant p-values (>0.05) indicate that a variable discriminates for BASS. The idea is to check which variables and BASS are dependent or not.

As an example, we show the obtained values coming from two indicators. The first one is the one followed in this paper, Block IX. Block XI is also demonstrated to complement the first indicator.

Tables 3 and 4 show the χ^2 values per each variable component.

Table 3. χ^2 values per use of time components.

Indicator/ Component	Block XI: Use of Time (Caretakers)	Children	GI	GII	GIII
χ^2 *p*-value	1.37×10^{-11}	4.27×10^{-26}	2.88×10^{-26}	1.83×10^{-22}	1.57×10^{-22}

Table 4. χ^2 values per B11 RelConv components.

Indicator/ Component	B11Rel Conv	R1.RelU ConvG20	R2.RelU ConvJ20	R3.RelU ConvG21
χ^2 *p*-value	1.37×10^{-11}	2.95×10^{-06}	1.58×10^{-17}	1.88×10^{-10}

All variables were significant, which means we were not able to discard any variable. Following these criteria, all variables were taken into consideration regarding accuracy, and we were not able to define any criteria.

Subsequenty, we used the methodology proposed in this paper, as presented in Section 2.6.2. The Use of Time variable was shown in Table 2 in Section 3.2.3, and B11RelConv was shown in Table 4.

As can be seen from Table 5, the B09 Use of Time (Caretakers) indicator is the chosen variable because it has the highest contribution ratio (0.3290).

Table 5. $R_k, \widetilde{\Pi}_{Lock}, E_k$ table for each B11 RelConv component.

Indicator/ Component	B11Rel Conv	R1.RelU ConvG20	R2.RelU ConvJ20	R3.RelU ConvG21
R_k	0.61	0.66	0.68	0.66
$\widetilde{\Pi}_{Lock}$	0.38	0.33	0.34	0.39
E_k	0.23	0.22	0.23	0.26

Nevertheless, the indicator B11RelConv was the selected variable because, in this case, R3.RelUconvG21 was the selected variable with the highest contribution percentage. As it is shown in Table 5.

These are 2 examples of the feature selection process. This was repeated for all blocks, and this phenomenon was repeated. Almost all *p*-values obtained using *p*-values were better.

Validating the TFSM

We compared classifications using the methodology explained in Section 2.6.4.

a. Numerical validation

The *p*-values of Lebart tests were calculated using two variables: the BASS variable and the class variable, as explained in Section 2.6.4.

The $S_{\wp}$ of the the classification using the variables selected in TFSM is 0.343, which is higher than the percentage that results without applying the TFSM, where $S_{\wp'} = 0.29$.

b. Graphical validation

The first validation that was performed was graphical. Figure 16 shows a map with all the new data-driven third-generation indicators after the clustering was completed. Figure 16 shows how territories belonging to the same class are far from each other. Moreover, social experts dislike this classification. We see that similar BASS are not located together. This fact is not helpful in terms of territorial clusters, as the main goal is to create a methodology to cluster similar BASS that are geographically close. Comparing this situation with the previous one (Figure 15), in Figure 16 it can be seen that the BASS are not as close as they should be.

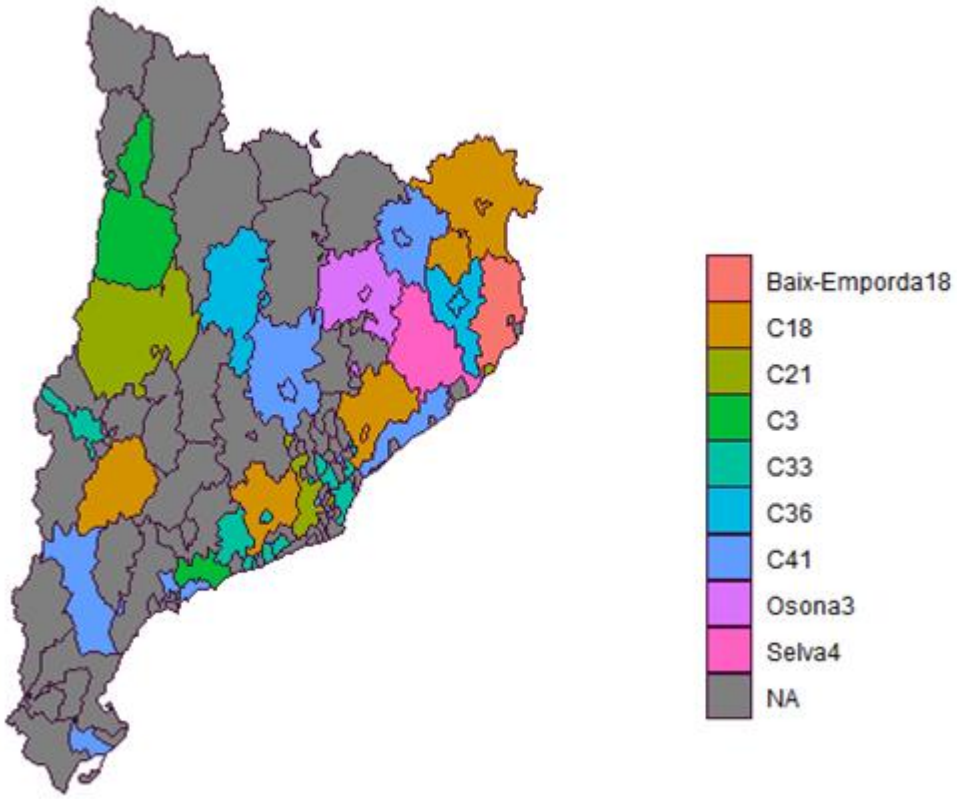

Figure 16. Territorial representation of clustering with Indicators generated.

4. Discussion, Conclusions, and Future Work

When data regarding several thematic blocks with different numbers of variables contain territorial references and a global analysis is required to find groups of territories, there is a need to find the best method to determine the representative variables of each block to be used in the global clustering process. Section 2.1.1. reviews a number of methods available for feature selection; most of the literature is oriented to feature selection in the supervised machine learning field. Nevertheless, our case is in the context of clustering, which is an unsupervised learning process, and an additional constraint requires attention since our data are territorial, and the territorial cohesion of the resulting clusters is also a goal. Otherwise, decision making might incur some applied inconsistencies. This paper provides several contributions to tackle several challenges associated with this main goal:

- Territorial feature selection method: This methodology is intended to build groups of territorial locations with territorial coherence based on clustering that might be interpreted through TLP based on the thermometer method and provide groups of BASS to be managed in a common way from the application domain point of view. This method is the main contribution proposed in this paper and includes the identification of the variable with the best performance in the global clustering (see details in Section 2.6.4).
- Thermometer: This is a new tool that assigns basic traffic light colors (green, yellow and red) to ranges of values for numerical variables or to the modalities of qualitative variables so that colors are associated with the semantics of the variable. It is a knowledge acquisition tool that allows domain experts to transfer semantics to the machine. It is formalized in this paper and enlarged with a fourth color (violet) used to represent missing values (see Section 2.6.1) and the DD2gI.
- TLP based on the thermometer method: This is a new method to automatically determine the color of each TLP cell using the knowledge and semantics formalized in a thermometer (see Section 2.6.2). It has been validated by experts based on the results of a real application on the INSESS-COVID19 database and by a comparison with

the traditional methods of building TLPs. This significantly increases the potential of the TLP tool, which was traditionally built by visual inspection of the conditional distributions of the variables with regard to a class variable.

- Data-driven second-generation indicators (DD2gI): This is a new methodology that enriches the data-driven third-generation variable creation method presented in [1] with the introduction of the thermometer method combined with clustering and traffic light panels. It is described in Section 2.6.3.
- Index of potential explainability: This is a new index based on the Lebart test values for qualitative variables computed versus location. It is used as the metric for selecting candidate variables inside TFSM. It is described in Step 4 of Section 2.6.4.

The set of contributions proposed in this paper was aligned to find groups of people from coherent territorial areas that can be treated together, and our proposed method was used to tackle real data regarding vulnerable populations and the INSESS-COVID19 database.

Decision making involving vulnerable people is sensitive and usually risky because a mistake might have terrible consequences for people already in fragility. The main goal of INSESS-COVID19 was originally to help social services in Catalonia support decisions for improving social services after the COVID-19 pandemic. The project started in May 2020, and in December 2020, a global report with the findings obtained from the INSESS—COVID19 database regarding all of Catalonia was published [3]. This report contained a large number of significant results and a very substantial view of what was happening with vulnerable populations in 20 daily aspects as a consequence of the first wave of lockdowns. The report allowed the government to make general decisions, such as opening some new services to attend to the emergent needs of the people or revising other services according to the results observed. This project had many impacts in the media, received awards, and helped the Catalan government make decisions right after the first pandemic lockdown, being useful in consolidating solutions that were enacted before DD2gI.

The target database comes from the answers of 971 vulnerable citizens who answere the INSESS-COVID19 questionnaire, which featured 195 questions from 19 daily topics. The preprocessed INSESS-COVID19 database presents up to 286 useful variables by including some new second- and third-generation variables that provide added value to improve the quantity and quality of information that the dataset can provide.

Further analyses of the INSESS-COVID19 dataset, including conditional analyses of specific territories, make it evident that other needs were not attended to because they were focused on a location and were involuntarily hidden under the global patterns found in the general analysis. This created the need to go further with a systematic approach to report locally to each territory (ideally at the BASS level). However, the sample was not sufficiently large to descend to this level of granularity without a high risk of violating statistical secrecy and the reidentification of individual participants. This raises the need to find a new approach to allow joint reports for groups of similar BASS so that the sample size guarantees participants' privacy while also considering territorial cohesion, which was a difficulty in the development of the paper.

The selection of the representative variables for each original thematic block of the survey to be inputted in the territorial clustering process requires new criteria, which were developed and presented in Section 2.6.4. The territorial feature selection method proposed in this paper is shown to provide a better grouping of BASS than previous works [1]. The comparison of the resulting clusters shows how the explainability of the clusters obtained with TFSM is higher than those obtained in [1] according to the S_P criterion. Additionally, the visualization of the obtained clusters over maps (Figures 15 and 16) shows more cohesive groups in the TFSM clustering, where several similar BASS are grouped together in a class, achieving one of the goals of the paper: clustering similar territories. The main clustering conclusions are the following. From the 11 clusters obtained, 5 represent a single BASS, and the other 6 groups represent between 4 and 11 BASS in larger groups with similar characteristics. The BASS clusters can therefore range from a single BASS (class Sant Joan Despí), which comprises people with good coexisting situations but many needs

and precarity regarding their position in the economy and their working conditions, to the Vilassar class, the members of which enjoy better conditions regarding coexistence, their living situation, the economy, and challenges regarding social subsidy access and participation in the community. This can be observed by a gradation of intermediate clusters where gradually, emotional support is less needed, economical support improves, or working conditions improve. The patterns elicited in this analysis are extremely useful to understand the special situation of immigrants (C31, distributed through 5 BASS with economic problems, working needs, and difficulties in participating in the community) or other patterns that could be connected with the activation of new specific protocols to attend to special vulnerabilities. Additionally, the results were published within a few days of data collection closing thanks to the powerful data science methodology designed in the INSESS-COVID system.

The contributions of this work were illustrated with the particular case of the INSESS-COVID19 dataset, but the methodology proposed is general and might be applied to any other kind of dataset (either coming from a survey or not) provided that the data relate to different thematic blocks (which nowadays is a very common situation). In fact, it has already been used in other real applications with successful results, such as the characterization of a non-profit association in Catalonia (Colla Castellera Jove de Barcelona) or the third-sector entities of three different pilot cities (Mataró (Spain), Prato (Italy) and Varazdin (Croatia)) within the European project SMP-COSME-2021-RESILIENCE 101074115-DIMCARE.

The limitations of this study arise because the territorial data perform poorly with traditional clustering methods in the sense that coherence in territorial groups are not guaranteed. Indeed, we needed for discovered groups to be neighborhood locations with similar characteristics, and that need was not trivial. Moreover, in the variable selection step, the strong association between all categorical variables added noise and confusion to attempts to properly identify the most discriminant variable in each view by traditional ranking criteria. Nevertheless, this proposal solved the main objective of the paper, which was to find territorial clusters from individual coherence. The TSFM is currently limited to qualitative variables, as in the target application, there were no numerical variables to be considered, although an extension of the proposal to numerical variables seems reasonable.

From a practical point of view, the INSESS-COVID19 dataset, including the conditional analyses of specific territories, demonstrated that some vulnerabilities of the population were not attended to on time because they were focused in a single location and were involuntarily hidden under the global patterns found in the general analysis. This created the need to go further with a systematic approach to report locally to each territory (ideally at the BASS level). However, the sample was not sufficiently large to descend to this level of granularity without a high risk of violating statistical secrecy and enabling the reidentification of individual participants. This raised the need to find a new approach to allow joint reports for groups of similar BASS so that the sample size can guarantee the preservation of participants' privacy, while territorial cohesion can also be considered. On the other hand, the data collection was originally designed from presential workshops in the territories, but the pandemics enlarged and precluded presential workshops. For this reason, we designed a new landing page and continued creating a new format for our workshop that included partial streaming sessions and online synchronous or asynchronous workshops.

In future works, further improvement on the automatic interpretation of the clusters will be in progress. On the one hand, regular expressions will be introduced to create textual descriptions of the clusters according to the results of the TLP and the profiling analysis. Additionally, the automatic sorting of rows and columns of TLPs will be tackled through the application of intelligent processes that can take into account the semantic relationship between thematic blocks. Currently, some other real applications are ongoing in the field of tourism and regarding an inspiring activity devoted to children that collected data on the electrical consumption of their houses and all the electrical devices used (TV, heating, etc.) so that they could understand which devices consumed more electricity, bring some recommendations home, and develop an interest in the data mining and AI fields for

future education. This last initiative was developed under the gender equality plan at UPC and the activities to reduce the gender gap in STEAM.

Author Contributions: Conceptualization and methodology, K.G. and X.A.; software development, K.G. and X.A.; validation, X.A. and K.G.; formal analysis, X.A.; investigation, X.A. and K.G.; resources, X.A.; data curation, led by K.G. and carried out by X.A.; writing—original draft preparation, X.A.; writing—review and editing, X.A. and K.G.; visualization, X.A. and K.G.; supervision, K.G.; project administration, K.G.; funding acquisition, K.G. and X.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Development Cooperation Centre (CCD) of the UPC under a special COVID-19 call (CCD-COVID-L019) and by Generalitat de Catalunya with the predoc grant no.: 2021-FISDU-00409.

Data Availability Statement: As the research uses the personal data of vulnerable people, the data are hosted in a UPC server compliant with the current RGPD, and individual data are not available from this project, as agreed with the respondents. Only aggregated data can be provided.

Acknowledgments: The authors want to thank their partner iSocial and especially Toni Codina for his contributions to the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Angerri, X.; Gibert, K. Preprocessing and Artificial Intelligence for increasing explainability in Mental Health. *Int. J. Artif. Intell. Tools* **2023**, *32*, 2. [CrossRef]
2. Gibert, K.; Angerri, X. The INSESS-COVID19 Project. Evaluating the impact of the COVID19 in social vulnerability while preserving privacy of participants from minority subpopulations. *Appl. Sci.* **2021**, *11*, 3110. [CrossRef]
3. Gibert, K.; Codina, T.; Angerri Torredeflot, X. *Informe INSESS-COVID19: Identificació de Necessitats Socials Emergents Com a Conseqüència de la COVID19 i Efecte Sobre els Serveis Socials del Territori*; Intelligence Data Science and Artificial Intelligence Research Center (IDEAI): Barcelona, Spain, 2020.
4. Sevilla-Villanueva, B.; Gibert, K.; Sànchez-Marrè, M. Identifying nutritional patterns through integrative multiview clustering. *Artif. Intell. Res. Dev.* **2015**, *277*, 185. [CrossRef]
5. Sevilla-Villanueva, B.; Gibert, K.; Sànchez-Marrè, M. A methodology to discover and understand complex patterns: Interpreted Integrative Multiview Clustering (I2MC). *Pattern Recognit. Lett.* **2017**, *93*, 85–94. [CrossRef]
6. Bickel, S.; Scheffer, T. Multi-view clustering. In Proceedings of the 4th IEEE International Conference on Data Mining, Brighton, UK, 1–4 November 2004; pp. 19–26.
7. Jundong, L.I.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
8. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. *Adv. Neural Inf. Process. Syst.* **2005**, *18*.
9. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
10. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef] [PubMed]
11. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
12. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
13. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 856–863.
14. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
15. Davis, J.C.; Sampson, R.J. *Statistics and Data Analysis in Geology*; Wiley: New York, NY, USA, 1986.
16. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
17. Jacob, L.; Obozinski, G.; Vert, J.-P. Group lasso with overlap and graph lasso. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 433–440. [CrossRef]
18. Núñez, H.; Sànchez-Marrè, M. Instance-based learning techniques of unsupervised feature weighting do not perform so badly! *ECAI* **2004**, *16*, 102.
19. Lebart, L.; Morineau, A.; Fénelon, J.P. *Traitement Statistique Des Données*; Dunod: Paris, France, 1990; p. 34.
20. Gibert, K.; Sevilla-Villanueva, B.; Sànchez-Marrè, M. The role of significance tests in consistent interpretation of nested partitions. *J. Comput. Appl. Math.* **2016**, *292*, 623–633. [CrossRef]

21. Gibert, K.; Sànchez–Marrè, M.; Izquierdo, J. A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Commun.* **2016**, *29*, 627–663. [CrossRef]
22. Torres, P.; Cruz, C.H.; Patiño, P.J. Índices de calidad de agua en fuentes superficiales utilizadas en la producción de agua para consumo humano: Una revisión crítica. *Rev. Ing. Univ. Medellín* **2009**, *8*, 79–94.
23. Vergara, C.; Arregui, I.; Balaguer, A.; Gómez, T.; Sandoval, C.; Sànchez Marrè, M.; Gibert, K. Learning on the relationships between respiratory disease and the use of traditional stoves in Bangladesh households. In Proceedings of the 8th International Congress on Environmental Modelling and Software Met, Toulouse, France, 10–14 July 2016.
24. Zhao-Hui, L.U.; Cai, C.-H.; Zhao, Y.-G.; Leng, Y.; Dong, Y. Normalization of correlated random variables in structural reliability analysis using fourth-moment transformation. *Struct. Saf.* **2020**, *82*, 101888.
25. Karina, G.I.O. The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Commun.* **1996**, *9*, 36–37. [CrossRef]
26. Ward, J.R.; Joe, H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
27. Glielmo, A.; Husic, B.E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised learning methods for molecular simulation data. *Chem. Rev.* **2021**, *121*, 9722–9758. [CrossRef]
28. Minghua, L.I.; Ferretti, M.; Ying, B.; Descamps, H.; Lee, E.; Dittmar, M.; Lee, J.S.; Whig, K.; Kamalia, B.; Dohnalová, L.; et al. Pharmacological activation of STING blocks SARS-CoV-2 infection. *Sci. Immunol.* **2021**, *6*, eabi9007.
29. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 86–97. [CrossRef]
30. Gibert, K.; Cortés García, C.U. Weighting quantitative and qualitative variables in clustering methods. *Mathw. Soft Comput.* **1997**, *4*, 1997.
31. Lefkovitch, L.P. Conditional clustering. *Biometrics* **1980**, *36*, 43–58. [CrossRef]
32. Gibert, K.; Garcia-Rudolph, A.; Garcia-Molina, A.; Roig-Rovira, T.; Bernabeu, M.; Tormos, J. Response to TBI-neurorehabilitation through an AI& Stats hybrid KDD methodology. *Med. Arch.* **2008**, *62*, 132–135.
33. Gibert, K.; Conti, D.; Vrecko, D. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *Environ. Eng. Manag. J.* **2012**, *11*, 931–944. [CrossRef]
34. Gibert, K.; Conti, D. aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. *AI Commun.* **2015**, *28*, 113–126. [CrossRef]
35. Gibert, K.; Conti, D.; Sànchez-Marrè, M. Decreasing uncertainty when interpreting profiles through the traffic lights panel. In *Advances in Computational Intelligence, Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, 9–13 July 2012*; Springer: Berlin/Heidelberg, Germnay, 2012; pp. 137–148. [CrossRef]
36. De Rham, C. La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Cah. De L'analyse Des Données* **1980**, *5*, 135–144.
37. COVID-19. Web del Project INSESS-COVID19. Available online: https://insess-covid19.upc.edu/ (accessed on 27 April 2023).
38. DIXIT Centre de Documentació de Serveis Socials. Available online: https://dixit.gencat.cat/ca/detalls/Noticies/tsf_presenta_eina_cribratge_ajudar_identificar_gestionar_casos_socials_complexos.html (accessed on 15 February 2021).
39. Pla Estratègic de Serveis Socials. Available online: https://treballiaferssocials.gencat.cat/web/.content/03ambits_tematics/15serveissocials/pla_estrategic_serveis_socials/Pla_estrategic_serveis_socials_catalunya_NOU/01_Plana_principal/1.-2020-12-29-Pla-estrategic-de-serveis-socials-2021-2024.pdf (accessed on 15 February 2021).
40. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]