Data Preparation for Longitudinal Data Mining: a case study on human ageing

Caio Eduardo Ribeiro[1], Luis Enrique Zárate[2]

Pontifícia Universidade Católica de Minas Gerais, Brazil
caioedurib@gmail.com
zarate@pucminas.br

Abstract. An adequate preparation of a database is essential to the extraction of useful knowledge contained in it. On longitudinal studies, that follow a fixed set of records through a time period, the data preparation process should adapt to the features added to the database by the temporal aspect of data. This article presents the data preparation process of a real longitudinal database, from a human ageing study. The process addresses the conceptual feature selection of the attributes in the database, and its pre-processing, including noisy data elimination, variable merging, discretization, outlier detection, and missing data analysis. The guidelines to the procedures were generalized, allowing their replication on other longitudinal databases, for similar studies.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining; H.2.m [**Miscellaneous**]:

Keywords: data mining, data preparation, data pre-processing, knowledge discovery, longitudinal databases, longitudinal data mining

1. INTRODUCTION

The process of knowledge discovery in databases (KDD) aims to discover useful non-trivial patterns through data mining algorithms. The data mining phase is preceded by a series of preparation steps, critical to make the found knowledge useful and correct [Fayyad et al. 1996].

There are several suggested database preparation steps in the literature, such as feature selection, the detection and elimination of inconsistencies and redundant data, missing data analysis and outlier detection, among other tasks. These usually are executed considering the characteristics of the database and the goals of the project, which directly influence how the data preparation techniques are applied. When applied appropriately, the data preparation reduces distortions, inconsistencies and polarization in data, apart from contributing to the performance of the data mining algorithms. All these actions collaborate with more valuable and reliable results on the knowledge discovery process [Pyle 1999].

Currently, it is possible to register data for extended periods of time, introducing a temporal aspect to it. Thus, we begin to handle historical data that must be explored through a temporal point of view, for its best understanding. Mining this temporal data allows the identification of cause-effect relations, and more accurate predictions due to the capability to observe the history and progression of the individual states of the data [Roddick and Spiliopoulou 2002]. It is recommended that the traditional KDD methodologies be restructured to attend the characteristics of temporal data mining [Last et al. 2001].

A class of temporal registers of data are the longitudinal studies, where the same sample of records

Copyright©2016 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

is followed through time, to characterize certain aspects of their evolution [Diggle et al. 2002]. The main difference between a longitudinal study and a regular temporal study is the analysis of the evolutionary behaviour of the sample, which is not expressed only in terms of seasonality, tendencies and averages, but also, and mainly, through approaches that ease the identification and analysis of phenomenons related to the passage of time. This can be achieved by comparing the data from a determined period, or wave, with its versions in distinct instants or time periods, always referring to the same records.

As stated previously, the domain problem characteristics and the type of database being analysed impact on the best way to perform the database preparation. When working with data that presents the longitudinal aspect, the traditional data preparation techniques need to be assessed in a new light. Moreover, to ensure that the longitudinal information of the data is kept, we need to make adaptations in the preparation strategies.

Within the context of longitudinal studies, an area that is gaining attention from the scientific community and governmental agencies is the human ageing study. The increasing attention to this area is due to the populational ageing phenomenon, expected for the coming decades. Studies indicate that the elderly population will surpass 21.5% of the worldwide total by 2050, a substantial increase in comparison to the current 12.3%. Such substantial growth will have great social and economic impacts [United Nations 2015]. Aiming to minimize the impact of populational ageing on the different social spheres, through the observation of the evolution of distinct environmental factors in the lives of sets of individuals, health researchers seek to formulate and test hypotheses of how the biological ageing is affected by individual choices and the environment on which a person is inserted.

Despite the growing importance of studying human ageing, our review to date found no work in which the KDD and data mining processes are applied to analyses of this global phenomenon. In order to encourage new longitudinal data mining (LDM) studies, in this article we discuss the different tasks of data preparation when applied to longitudinal databases (LDBs). As a case study, and foundation to our generalizations, we consider a real database used in human ageing studies.

This article is organized as follows: the second section presents concepts about human ageing and on longitudinal studies through a data mining perspective. All the data preparation method of the case study is presented on the third section, whilst detailing and generalizing the procedures adopted, to similar projects. Finally, the fourth section has our conclusions about this work.

2. BACKGROUND

2.1 Human ageing studies

Because of the gradual ageing of the world population, currently there is a greater interest in understanding how genetics and environmental factors determine the way age affects people's cognitive functioning, health and psychological state. The populational ageing impacts the entire structure of society, specially regarding social security issues, because the ratio of working versus retired people will decline, which will have severe social and economic implications [Lutz et al. 2008].

The studies on human ageing that focus on environment conditions (which are easier to handle than the genetic conditions) aim to determine the way personal choices, and the inherent characteristics to the environment in which the individual lives, affect one's biological ageing. In order to understand this phenomenon, an interdisciplinary evaluation of the several aspects that compose the environment is necessary. Efforts at the national level are being carried out, mainly in nations where populational ageing is more aggravated, such as censuses to collect data on the ageing of their citizens. As mentioned, longitudinal studies follow a set sample of individuals through years, and generate databases with detailed descriptions of various aspects of their lives.

The most common form of longitudinal data analysis found in the literature is the usage of classical

statistical techniques. For instance, it is usual to find regression analysis studies, which aim to infer the value of a dependent variable by studying a set of independent variables [Cacioppo et al. 2006]. Hypotheses tests, and investigations of correlations between variables can also be found often. As an example, on the study of Kim et al. [2012], the authors tried to find correlations between self-reported health and self-reported life satisfaction on a set of individuals.

Among the large scale longitudinal studies, the English Longitudinal Study of Ageing, ELSA, is one of the most prominent [Marmot et al. 2015]. The ELSA study has, in each of its 6 waves, thousands of respondents from United Kingdom households inhabitants, that are visited and interviewed every two years (duration of a wave of the study). In its current form, the ELSA began officially in 2002, and the discussed features include demographic, economic, social, of physical, mental, and psychological health, and cognitive functions.

Another longitudinal study of human ageing worth mentioning is the Survey of Health, Ageing and Retirement in Europe, SHARE, a continental level study that unites the efforts of twenty European countries, in addition to Israel. Being the most geographically distributed study of the world, SHARE has the largest sample of records, an important feature to some studies. However, the comprehensiveness of the ELSA was considered more suitable for our study, because it addresses several physical health aspects through test results (such as psycho-motor tests, cognitive tests, memory tests and blood tests), performed by health professionals who participate in the visits to the respondents of the study. For that reason, the ELSA database has been chosen for the case study in this work.

The ELSA is intended for persons 50 years of age or older, in order to follow the participants for years prior to their retirement and beyond, which allows a detailed analysis on the evolution of the observed aspects [Banks 2006]. Most ELSA questions have predefined response options, making the generated database predominantly categorical. The ELSA database had some of its attributes and records altered throughout the study, due to updates in the samples and questionnaires, which made a pre-processing of the data necessary, before the longitudinal analysis itself. For this study, we used only the records and attributes common to all ELSA waves, creating a longitudinal version of the database

2.2 Longitudinal Studies and Data Mining

With the increasing interest on studying human ageing, large scale longitudinal studies have been initiated in several countries, producing databases made available for use in scientific researches. As stated previously, our reviews found no published works using data mining techniques on longitudinal studies databases. However, we believe that a more comprehensive analysis of the frequently complex problems addressed on social studies of human ageing, would be made easier and more efficient through data mining techniques.

Formally, a longitudinal database is a temporal database with the same identity to all the time units. We define as longitudinal any database that can be described as a matrix M, composed by the Cartesian product of three vectors: r (of records), a (of attributes), and t (of time units), as shown in Equation 1. The representativeness of a database is related to its dimensionality, which is affected by the size of each of the vectors.

$$M_{rat} = [r] \times [a] \times [t] \tag{1}$$

As the same sample of records is followed through time in a longitudinal study (the records have unique identifications, that can be used to trace the same record on different waves), the attributes of the database have a temporal aspect, and their values present previous and posterior states, considering the different waves of the database.

It is necessary to consider the temporal information on data during the planning and execution of data preparation tasks, whilst the data mining algorithms must be adapted to handle this characteristic. LDBs inherently have greater data volume and complexity, and the focus of LDM studies is usually to identify the causal relation of an observed effect, and the evolution of the effects to a set of attributes on the database [Last et al. 2001].

Data mining applied to longitudinal studies brings forth the possibility to discover and interpret patterns as: a) cohort effects, based on previous causes (effects that are specific to a sample of records); b) seasonal longitudinal patterns (behavior that repeat in determined intervals of time, or happen in function of temporal events); or c) effects of the passage of time (evolution of the observed aspects). The traditional data mining algorithms of rule association, classification, and clustering must be adapted to achieve these goals.

3. DATA PREPARATION FOR LONGITUDINAL DATABASES

As stated in Pyle [1999], many data miners neglect the data preparation tasks prior to the execution of mining algorithms, and that behavior can cause the failure of a KDD project. The time spent on data preparation is a necessary investment to assure that the discovered knowledge is relevant. In practice, experience shows that the time need to fully prepare a database can reach 75% of the total time spent in a KDD project [Jermyn et al. 1999].

The data preparation process aims to assure that the data is as relevant as possible to the KDD project. We consider relevant data that represents reality, and is presented in a sufficient quantity of records, or instances, to ensure that the found results are applicable [Kotsiantis et al. 2006].

When applying KDD to LDBs, we should not consider temporal data as a simple collection of unordered events, ignoring the fact that these data reflect values referring to the same set of records, with a chronological order, which impacts on the posterior waves values of the data [Shahnawaz et al. 2011]. In this section, we describe the actions performed during the data preparation process for the ELSA database, on a LDM project. These actions are detailed to highlight the differences that the characteristics of this kind of database incur on the preparation process, moreover, the actions generalized to allow their replication on similar databases. In this article, we consider as parts of the data preparation process all the actions performed since the definition of the goals of the study, reaching the point of executing the data mining algorithms.

There are some intrinsic characteristics to LDBs, due to the way longitudinal studies are conducted:

(a) As the same person has to be located each wave to respond to the ELSA questionnaires, the amount of missing data and discontinued records can become large. Furthermore, the studies themselves evolve over time, with changes on their approach and goals that change data structures, such as the emergence of new variables and the elimination of others, as a result of new guidelines and changes in the scope of the study;
(b) A typical feature of the longitudinal studies is their comprehensiveness, so their databases usually have large dimensionality (large amounts of records and attributes). That hampers the manipulation and processing of these databases by algorithms, hence the need to a rigorous feature selection and preparation process.

With these LDB aspects in mind, the data preparation process discussed in this article is grounded in two phases: 1) definition of the problem and conceptual feature selection; and 2) data pre-processing. Initially, in the first phase, we propose a previous study on the domain problem, in order to establish the most relevant attributes in the database, eliminating those of lesser relevance through a process we referred to as a conceptual feature selection. The second phase of the process consists in guaranteeing that the database is in the correct format, identifying and eliminating inconsistent, imprecise or missing data from the database, which undermine the performance of mining algorithms. The goal of the

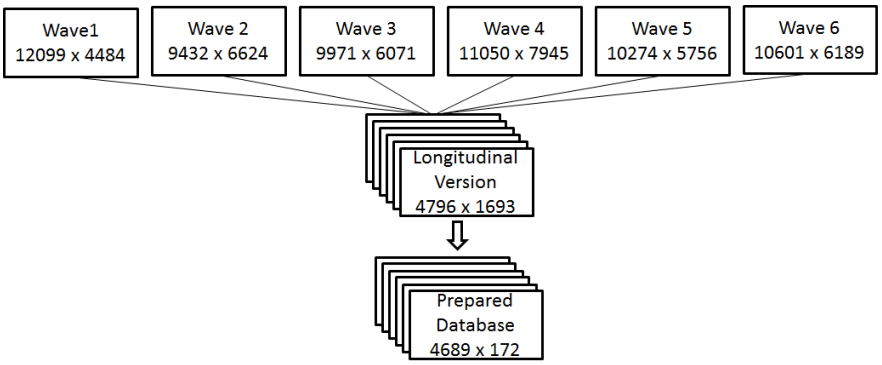

Fig. 1. Dimensions of the different versions of the ELSA database.

preparation phase is to ensure that only relevant data remains represented, and that their representativeness is enough for the mining algorithms to discover the patterns that represent the knowledge existing in the database. The Figure 1 exhibits the original dimensions (records x attributes) of the 6 waves of the considered study database, in addition to the dimensions of the longitudinal and final versions of that database, after the preparation process. The resulting database has 4,689 records and 172 attributes, in each of its 6 waves (referring to the 2002-2012 period) considered on this study.

3.1 Previous Study and Conceptual Feature Selection

3.1.1 *Problem Definition.* The first step of the first phase of the data preparation consists of defining the domain problem that will be explored on the study. Through this definition, it is possible to establish the scope of the project, and which variables (attributes of the database) are relevant to the discovery and analysis of patterns. For example, in the case study, we research the influence of the environment on the process of human ageing. To prepare an adequate database, it was necessary to study the composition of the environment, and how the variables relate and were collected.

A previous study [Ribeiro and Zárate 2014], aimed to conceptually model the problem of human ageing to highlight the environmental variables considered most relevant. This definition of the relevance of each attribute supports the feature selection process. Therefore, the selection of the environmental variables that were most frequently found in studies on human ageing, made in the aforementioned work, guided the step of conceptual feature selections, described hereafter. Figure 2 contains a selection of the most addressed environmental aspects, organized in: economic, social, life-quality, physical health, and mental health categories.

The goal of a KDD project, which a database has been prepared to, is discovering effects of the passage of time on the response of questions about the several environmental aspects considered in the study. Thus, as the scope and goals of the project have been defined, the data preparation process can be initiated.

3.1.2 *Database composition.* A LDB corresponds to a Cartesian product of three vectors (as shown in Equation 1), which implies that the same records and attributes repeat over all waves of the database, represented by the time vector t. It is expect that, as the longitudinal studies evolve and previous results are found, the attributes of their databases modify to comply with new demands of the study, besides the addition of new records (respondents included in the study), and removal of some of the records (respondents that do not participate on posterior waves). Such changes prevent the longitudinal analysis of the database, because it is impossible to track the changes of values in the database. Therefore, for this study, each register or attribute can only be kept in the database if it is present in all waves of the study, as defined in Equations 2 and 3.

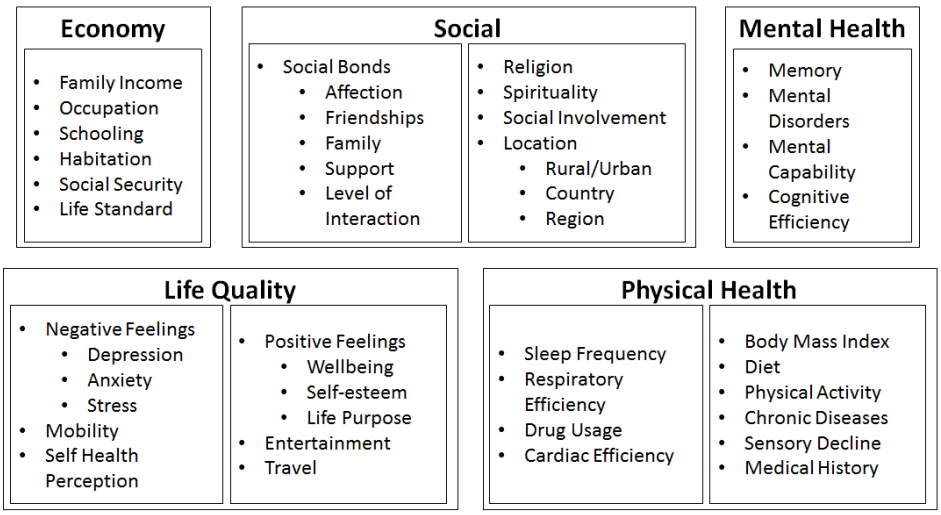

Fig. 2. Most addressed environmental aspects in published works.

$$r_i \in M_{rat} \mid r_i \in M_{rat_j} \forall j \tag{2}$$

$$a_i \in M_{rat} \mid a_i \in M_{rat_j} \forall j \tag{3}$$

As mentioned, the ELSA is a national level study, addressing several environmental aspects of the life of its respondents. The resulting database differs for each wave. Thus, to allow a longitudinal analysis of the database, we removed from it the attributes and records that were not present in each of the 6 considered waves (2002-2012). This observation reduced the database to a total of 4,796 records and 1,693 attributes.

One of the main features of the LDB is the large volume of data produced by the temporal axis, which increases the importance and necessary care in the conceptual feature selection and data filtering tasks. In a longitudinal study, ensuring that the knowledge is represented by a minimal quantity of data that ensures representability and comprehensibility is not only a matter of avoiding useless or redundant knowledge found. These tasks are directly related to the viability of the algorithmic processing of data, to the interpretability of the knowledge and, thus, to the success of the KDD project [Paes et al. 2013].

3.1.3 *Conceptual Feature Selection.* The conceptual feature selection aims to keep on the database only the attribute that are most relevant to the execution of the study. It would be worthwhile to note that an attribute might be considered relevant for different reasons. For instance, being related to the objectives of the research, having relevance because it improves the distribution of the records, or the precision of a classification algorithm [Blum and Langley 1997].

Selecting attributes implies in removing from the database the less relevant ones, reducing the complexity of executing mining algorithms with the data. Knowledge on the problem domain of the study is key to correctly defining relevance, because the attributes are judged according to the understanding of their relationship to the problem addressed.

Some attributes become more relevant when presented with a temporal aspect. To these attributes, their static version carries little significance, however a series of values adds important information. For example, a measure of heart frequency carries more information if there is a history of measures to that patient. Another example would be the psychological history of a patient, used to identify high levels of stress.

On the conceptual feature selection phase, to make the relevant attributes choice as adherent as possible, firstly it was necessary to define what composes the environmental influence on human ageing studies, object of study in this project (see Figure 2 to check the aspects considered). Through the knowledge obtained on the previous study and the definitions of the conceptual modelling of the problem, an individual analysis of the questions present in the questionnaire elaborated to the longitudinal study. We discarded the attributes that were not related to the aspects identified on Figure 2, as the aspects indicated in the model are considered the most important on ageing study. From the 1,693 attributes in the longitudinal version of the ELSA database, only 275 have been kept after this task, considered the most relevant to the study. It is worthwhile to note that it is highly recommended to apply this conceptual feature selection process prior to more elaborated computational and statistic procedures, such as wrappers and filters algorithms. These techniques should be applied after the data preparation process, and before using mining algorithms, providing a formally based dimensionality reduction. In contrast, the conceptual feature selection is based on explicit and tacit knowledge of the domain problem, and can be easily applied even to bulky databases.

Databases are composed by facts and judgments, the last ones being assumptions made to come closer to the reality of the problem. Therefore, after the conceptual feature selection is finished, we suggest a second analysis on the selected attributes. It is important to have a control of the level of facts and judgments on the attributes in a database. If a database is composed only by facts, it is impossible to discover unprecedented knowledge, and as judgments are added the exploration capacity increases, but so does the distance to the real characterization of the problem. Finding this balance can be the difference between discovering obvious or relevant knowledge at the end of a KDD process.

Thus, if a database has low factual level, it is recommended that some attributes with high judgment level be eliminated, reducing the dimensionality from the database, while also enriching its capacity to generate useful unprecedented knowledge. The decision to eliminate an attribute depends on knowledge on the information it represents. For example, in the case study, we eliminated an attribute that represented the question “How many employees does the company you work for have?”, and kept the question “How many hours a week do you stay at work, including extra and interval hours?”. The eliminated question adds information on the size of the company where the respondents work, however this information is not directly related to their economic situation. The kept question, however, represents an important information on the routine of the respondents, thus being more relevant to the study. Using this elimination by judgment level filters the database, using the previous knowledge on the relevance of the attributes.

The selection by judgment level considers the explicit knowledge obtained during the previous study phase, to discard attributes that are poorly factual, thus reducing the volume of the database. It is important that, in such cases, a careful study of the domain problem is used to minimize the damage that the lack of tacit knowledge from an domain expert brings onto the KDD project. Ideally, all the feature selection process should be validated by a domain expert, who evaluates the relevance of each attribute available to the study.

The intervention of the domain expert has been identified as a key aspect to success in knowledge discovery processes. This paradigm is named D3M (Domain-Driven Data Mining), and recommends the generation of methodologies to KDD processes that are oriented to the problem [Cao 2010]. According to D3M concepts, each task on the KDD process must be accompanied and validated by the domain expert, which makes the process more adherent.

3.2 Data Pre-processing

In this article, we consider as parts of the pre-processing phase all the procedures adopted to refine the database, before applying the data mining algorithms. These procedures are described hereafter.

3.2.1 *Noisy Data Elimination.* Noisy data arise from data collection or database generation, which are common in large data volumes. These data add false information that will be interpreted by the mining algorithms, and can lead to invalid results. Thus, the following actions should be performed to identify and eliminate noisy data:

(a) **Duplicate records analysis**: On longitudinal databases, such as the one in this case study, similar records on the same wave or distinct waves could be considered valid. These records can represent individuals with the same responses on the questionnaire, or an individual that kept his response through several waves.

(b) **Inconsistencies analysis**: A longitudinal database can, specially if the data is collected manually, present inconsistencies which are data registration errors that make the value of an attribute invalid. A conformity analysis on the values found in the database, to check if they are according to the formal definition of the expected values in an attribute should be done during the preprocessing phase. It is possible to infer the correct values of inconsistent data, but we must consider that assuming a value for an attribute can actually add more noisy data to the database, and harm the knowledge represented. A possible inconsistency example would be a negative value as a response to a question about the remuneration of a respondent.

(c) **Attribute transformations**: Some recoding and transformations on attributes can be necessary to make the database consistent. If an attribute has its possible values (options of answer) modified through the study waves, the values from the version with higher cardinality should be remapped to get consistent to the one with lower cardinality. This way, it is possible to keep coherence and consistency on the attributes, which is necessary for longitudinal analysis, with a minimal information loss.

Importantly, the ELSA database underwent a preliminary review by those responsible for study, eliminating possible insertion errors, duplicate records and inconsistencies. Still, some adaptations were necessary on the values of attributes referring to questions that had been somehow modified through the waves of the study. The attributes were recoded and their values remapped, so that the database had an unique and coherent version of the attribute, enabling its longitudinal analysis.

3.2.2 *Missing Data Analysis.* The missing data analysis seeks to adequately treat data with missing values, and dealing with them by eliminating the record, the attribute, or imputing a most likely value to the missing data. Different techniques to infer missing values on data can be applied, such as retrieving a likely value from other sources, in addition to the possibility of using calculations to determine a statistically more approximate value for the missing data (averages, medians, modes, among others) [Rubin 2006] [Silva and Zárate 2014].

However, in a longitudinal database, it is possible to resort to values from different waves to infer missing data with more reliability. If there is on the database a value for the attribute on the previous or posterior waves, we are more likely to being able to correctly infer the value to a wave where the value is missing. This characteristic of the longitudinal database makes imputing missing values more efficient, for using information directly related to the missing data to infer its value. This procedure requires tacit knowledge on the domain problem to evaluate its applicability.

In some ELSA questions, if the response is unchanged on subsequent waves to the first wave the question was asked, a code is registered on the database to indicate that the initial response was kept. In these cases, the real value of the attribute was recovered by retrieving the initial response to that question. Besides that, to each ELSA record a weight value is calculated, according to the amount of information that record adds to the database, meaning that individuals that responded a very low amount of questions got a smaller weight value. The records that received zero weight value on this evaluation were eliminated from the database.

3.2.3 *Feature Selection through the Amount of Information.* If an attribute has excessively low information, including it on the study might confound the interpretation of results, making the knowledge less understandable. On these cases, it might be more adequate to reduce the dimensionality of the database by removing these attributes with low information. By calculating the entropy (S) of an attribute, it is possible to ascertain the amount of information they add to the database, according to their variability [Li and Wang 2002]. The entropy calculation utilized considers the probabilities of the j possible values of an attribute, as described in Equation 4.

$$S = \sum_{i=1}^{j} P(i) \times \log_2 \frac{1}{P(i)} \tag{4}$$

The ELSA attributes whose entropy calculation resulted in a value too close to zero (the threshold defined was $S < 0.1$) were individually analyzed, instead of readily eliminated. In some questions, such as those that registered the existence of health conditions on the respondent, we expected a low entropy value, because the majority of the respondents did not suffer from those conditions. On the cases, the low variability was expected, and the information contained in attributes was considered important enough for them to be kept on the database. However on attributes where the low entropy found was considered an atypical behavior for the information represented, they were eliminated from the database.

3.2.4 *Outlier Detection.* Some records have values that are not consistent with the sample of the study, for being too discrepant. The elimination of these outliers can be done automatically trough algorithms that identify the discrepant records examining the database. Approaches using neural networks [Williams et al. 2002], filter algorithms [Liu et al. 2004], clustering analysis, among others, aim to increase the efficiency of outlier detection, an important task to increase the precision of the KDD process.

On the literature, it is recommended to readily eliminate outliers to keep the data representativeness from being distorted. However, on longitudinal studies, the outlier analysis must consider the temporal information represented by these records, which might be relevant to predict future states of the database. Thus, the decision of eliminating outliers or not gets harder, once valuable information for prediction might be found on these records.

Figure 3 shows the possible evolutions of a clustering case with outliers, over time. The initial situation can develop on the three ways shown, according to the strength of influence on the clusters and on the outliers. The first hypothesis is the adaptation of outliers (1), where the records modify their characteristics over time to fit the established patterns of existing clusters. On the second, adaptation of clusters (2), observed mainly on social behavior changes, the outlier influence makes the cluster behavior adapt gradually, modifying the characteristic values of that set of records (which can explain a change of patterns over time). On the third possibility presented, the outliers gather to form new clusters (3), modifying the study scenario. Therefore, on the temporal context, the outlier analysis can help explain and predict the changes on clusters over the time. These records, often discarded as noisy data, thus become valuable to longitudinal studies.

Regarding the outlier analysis on ELSA, we considered as outliers of the study only records which did not characterize the target audience of the study, of individuals that were at least 50 years old. Younger respondents had been included on the study because they met certain prerequisites, such as the prospect of reaching the minimum age for the study. The records with values that were discrepant from averages and modes on the database were kept, for future cluster evolution evaluations.

3.2.5 *Discretization.* The next task of the pre-processing phase consists in making the necessary transformations on attributes to enable the use of the chosen data mining tools. Some mining algo-

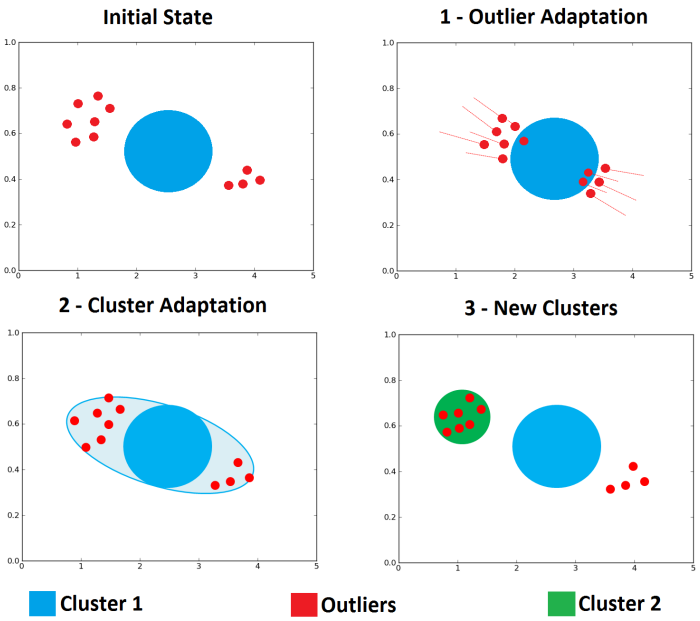

Fig. 3. Temporal outlier analysis.

rithms require that input data are categorical, creating the need to discretize the numerical attributes present in the database.

The discretization problem is not trivial, because of the number of different combinations that can be done by modifying the number and size of value bands created to represent categories on a continuous interval. Therefore, we can use heuristics that suit specific situations and seek an approximation of the optimal discretization. Typically, the choice of the most suitable technique is made in accordance with the distribution of values, desired number of intervals and the information represented by the attribute [García et al. 2013].

On the case study, the numerical attributes were individually assessed, and discretized according to the guidelines of *EqualFrequency* and *EqualWidth* discretization techniques [Li and Wang 2002], both widely used. The decision between the discretization techniques was based on the data distribution, and also on the information represented by the attribute. On the continuous attributes (economical variables, such as income) and quantification attributes (for example, number of children, and how many rooms on the residence), the *EqualFrequency* discretization was adopted, because the values of these attributes were not equally distributed. However, on questions referring to age or dates (for example, age when diagnosed with some health condition) and percentage questions (such as the well-being questionnaire questions, that should be answered with a number from 0 to 100), the *EqualWidth* discretization was used, because the data distribution was more uniform on those questions.

Understanding the problem and its attributes to the point of being able to use a discretization based on tacit knowledge can make the database more robust. The numerical attributes on ELSA were discretized using the described guidelines, following the precepts of both algorithms. On the Figure 4 four discretizations made in the case study are presented. The attributes discretized with the *EqualWidth* technique present values that represent upper and lower limits, and bands of the same magnitude between these. On the other hand, the attributes discretized with the *EqualFrequency* attributes do not have a visible pattern on their bands, because they were created considering the distribution of values in the database.

Percentual question: EqualWidth	
What are the chances that at some point in the future you will not have enough financial resources to meet your needs?	
-1	No answer
a	0
b	1-20
c	21-40
d	41-60
e	61-80
f	81-99
g	100

Age question: EqualWidth	
Approximately how old were you when first told by a doctor that you had a stroke? Years	
a	Never
b	Less than 30
c	30-39
d	40-49
e	50-59
f	60-69
g	70 or more

Quantity question: EqualFrequency	
How many living brothers or sisters do you have?	
-1	No answer
a	0
b	1-2
c	3+

Value question: EqualFrequency	
How much do you (and your spouse) owe to friends, relative or other private individuals?	
-1	No answer
a	0 - 999
b	1000-4999
c	5000 or more

Fig. 4. ELSA attribute discretization.

Question A: Are you often troubled with pain? A1: Yes. A2: No.	Question (A X B): Are you often troubled with pain? If you are, how bad is it most of the time? Is it... ? (A1&B1): Yes, Mild. (A1&B2): Yes, Moderate. (A1&B3): Yes, Severe . (A2&B1) || (A2&B2) || (A2&B3): No.
Question B: How bad is the pain most of the time? Is it... B1: Mild. B2: Moderate. B3: Severe .	

Fig. 5. Example of attribute merging.

3.2.6 *Variable Merging.* On a categorical database, it is possible to take advantage from this characteristic, by making a merge of highly dependent attributes in order to reduce the volume of data. On the case study, some attributes could be represented by a single variable, without great loss of information. Therefore, concluding the preprocessing phase and the data preparation of the ELSA database, 22 highly dependent attributes were merged on this procedure, creating 9 new attributes to replace them.

Consider an attribute A with cardinality $|A|$ and an attribute B with cardinality $|B|$. The Cartesian product of both attributes will have its cardinality as defined in the corollary shown in Equation 5. To merge attributes, firstly the Cartesian product of both categorical attributes is done, generating a new question with options of answer referring to all the possible combinations of answers to the merged attributes. Then, this new attribute is adapted, through a discretization that unites answers with a similar meaning. This, in order to keep the merges with a reduced number of options of answer, a desirable feature for clustering algorithms.

$$|A \times B| = |A| \times |B| \tag{5}$$

Figure 5 has an example of variable merging with two ELSA questions. The Cartesian products of the attributes has six options of answer, however three of them could be represented by a single response.

Table I exhibits a synthesis of the recommended guidelines on each task of the data preparation process, according to the procedures adopted in the case study. These actions can be replicated on data preparation processes for longitudinal databases with characteristics similar to the ELSA database. It is important to reinforce that the application of all these guidelines should be oriented by the objectives of the research being developed.

Table I. Guidelines for the preparation process.

Task	Objectives	Guidelines
Problem Definition	Establish the scope of the project and its goals	Conceptually model the problem through an exploratory study and/or the help of a problem domain specialist
Database Composition	Ensure that the database is longitudinal	Eliminate records and attributes that do not exist in every wave of the database
Conceptual Feature Selection	Select the attributes in the database that are relevant to the study	Evaluate the relevance of the variable, according to the previous study. Eliminate those which are irrelevant, and those which present very high judgment levels (are less factual)
Noisy Data Elimination	Identify and eliminate noisy data from the database	a) Only records with the same ID are considered duplicate. In a longitudinal database, resemblant records can correspond to similar responses of different respondents or waves. b) If the set of possible options of answer of an attribute changes, set a version with fewer options of answer as default, and adequate the others to it, remapping the answers on the database.
Missing Data Analysis	Recover the missing values of attributes	The previous and later values of an attribute are the most reliable way to infer its correct missing value.
Feature Selection through the Amount of Information	Eliminate attributes that add little information to the database	Attributes with low entropy should not be readily eliminated from the database. Establish a threshold value to the entropy (e.g. $S < 0.1$) and examine further the attributes that have small entropy values. The attribute might be of value to the database even if it has low variability, according to the information it represents.
Outlier Detection	Identify records that differ from common behavior found in the database	Outliers should not be readily eliminated from a longitudinal database. The outliers can be used in the longitudinal analysis, to understand the changes in the behavior of clusters of records that happen over time.
Discretization	Categorize numerical attributes, for use in mining algorithms that demand categorical variables	Examine the data distribution of the attribute. If it is poorly distributed, use *EqualFrequency* discretization, otherwise use *EqualWidth* discretization.
Variable Merging	Reduce dimensionality, representing highly dependent variables through a single variable	Create a new variable through the Cartesian product of the dependent variables. Discretize this new variable, if possible, reducing its number of possible values.

4. CONCLUSIONS

In this article, we presented a case study on the preparation of a longitudinal database originated from the English Longitudinal Study of Ageing. The adopted procedures were generalized, being able to be replicated on similar projects. Firstly, we defined the scope and objectives of the study, and these definitions guided the decisions made during the data preparation process. Next, through explicit knowledge obtained on an extensive domain problem study, we performed a conceptual feature selection on the database, aiming to choose the most relevant attributes. Finally, the database has gone through a pre-processing that eliminated noisy data and attributes with low information, recovered missing values, analyzed outliers, discretized numerical attributes, and merged dependent variables.

It is worthwhile to note that the conceptual feature selection, guided by a previous study on the problem domain and the analysis from a domain expert, is essential to the success of the LDM process. The relevance to the explored problem of the attributes that compose the database enriches the discovered knowledge, besides this selection resulting on a reduction of dimensionality that facilitates the computational processing of data, and the understandability of the extracted knowledge.

At the end of the data preparation process, the LDB suffered a reduction of about 90% of its data volume, with a final dimension of $r = 4,689$ records, $a = 172$ attributes, and $t = 6$ waves.

All the records kept on the database represent respondents that characterize the target audience of the study (age equal or superior to 50), and had a considerable weight value. From the remaining attributes, after the eliminations by conceptual feature selection, filtering by amount of information, and the variable merging procedure, the five categories of environmental aspects have the following representation: 30 economic variables, 47 social, 47 life quality, 33 physical health, 9 mental health and 6 identification variables that do not fit any category (unique record ID, gender, weight value, age, ethnicity and country of birth).

As future work on our case study, we recommend applying clustering mining algorithm on the database, to find patterns of respondents regarding information from all the different dimensions. The preprocessed database improves the mining results, providing data that is cleaner and more related to the investigated problem domain. In addition to that, its smaller volume will greatly reduce the execution time of the algorithmic analysis, usually of high complexity.Because the ELSA database refers to a questionnaire, there is an implicit tendency to neutral answers on opinion questions, that makes similarity calculations between different registers tend to the same average value. That issue, allied to the dimensionality problem, stated previously, complicates the process of separating the registers in groups, and the data mining process needs to address them. There are a lot of approaches that could treat this kind of problem, such as density based algorithms, hybrid clustering algorithms (that don't need the database to have only categorical or numerical values), and specific similarity and distance measures.

The techniques adopted on the data preparation process follow traditional precepts in the literature, however we discuss the special features of these techniques when used on longitudinal databases, that have a temporal aspect and serve a distinct purpose than traditional and common temporal databases. The difference on the database preparation is mainly due to the longitudinal information of the data, in other words, the fact that the values on the database have previous and posterior versions. On the preparation of longitudinal databases, the recovering of missing values becomes more precise, the outlier study can facilitate prediction of future tendencies on the database, and the feature selection has a larger impact on the success of the project. Regarding the KDD process as a whole, the evolutionary characteristic of the longitudinal database, of following a sample of records over time, brings in new possibilities and challenges that need to be explored and studied more profoundly.

Acknowledgement

The data were made available through the UK Data Archive. ELSA was developed by a team of researchers based at the NatCen Social Research, University College London and the Institute for Fiscal Studies. The data were collected by NatCen Social Research. The funding is provided by the National Institute of Aging in the United States, and a consortium of UK government departments co-ordinated by the Office for National Statistics. The developers and founders of ELSA and the Archive do not bear any responsibility for the analysis or interpretations presented here.

This work was conducted during a scholarship supported by the International Cooperation Program CAPES/COFECUB at the PUC-Minas University. Financed by CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

REFERENCES

Banks, J. *Retirement, Health and Relationships of the Older Population in England: The 2004 English Longitudinal Study of Ageing (wave 2).* Institute for Fiscal Studies (Great Britain), 2006.

Blum, A. L. and Langley, P. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence* 97 (1-2): 245–271, 1997.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., and A., T. R. Loneliness as a Specific Risk Factor for Depressive Symptoms: cross-Sectional and Longitudinal Analyses. *Psycology and Ageing* 21 (1): 140–151, 2006.

Cao, L. Domain-Driven Data Mining: challenges and Prospects. *IEEE Transactions on Knowledge and Data Engineering* 22 (6): 755–769, 2010.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: an Overview. In *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 1–34, 1996.

García, S., Luengo, J., Sáez, J. A., López, V., and Herrera, F. A Survey of Discretization Techniques: taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 734–750, 2013.

Jermyn, P., Dixon, M., and Read, B. J. Preparing Clean Views of Data for Data Mining. *Proceedings of the ERCIM Workshop on Database Research*, 1999.

Kim, S., Sargent-Cox, K. A., French, D. J., Kendig, H., and Anstey, K. J. Cross-national Insights into the Relationship between Wealth and Wellbeing: a Comparison between Australia, the United States of America and South Korea. *Ageing & Society* 32 (1): 41–59, 2012.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. Data Preprocessing for Supervised Learning. *International Journal of Computer Science* 1 (2): 111–117, 2006.

Last, M., Klein, Y., and Kandel, A. Knowledge Discovery in Time Series Databases. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 31 (1): 160–169, 2001.

Li, R.-P. and Wang, Z.-O. An Entropy-based Discretization Method for Classification Rules with Inconsistency Checking. In *Proceedings of the International Conference on Machine Learning and Cybernetics*. Vol. 1. pp. 243–246, 2002.

Liu, H., Shah, S., and Jiang, W. On-line Outlier Detection and Data Cleaning. *Computers & Chemical Engineering* 28 (9): 1635–1647, 2004.

Lutz, W., Warren, S., and Sergei, S. The Coming Acceleration of Global Population Ageing. *Nature* 451 (7), 2008.

Marmot, M., Oldfield, Z., Clemens, S., Blake, M., Phelps, A., Nazroo, J., Steptoe, A., Rogers, N., and Banks, J. English Longitudinal Study of Ageing: Waves 0-6, 1998-2013, 2015.

Paes, B. C., Plastino, A., and Freitas, A. A. Seleção de Atributos Aplicada à Classificação Hierárquica. In *Proceedings of the Symposium on Knowledge Discovery, Mining and Learning*. São Carlos, SP, Brazil, pp. 1–8, 2013.

Pyle, D. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

Ribeiro, C. E. and Zárate, L. E. Uma Revisão para Identificar Variáveis Ambientais que Influenciam o Envelhecimento Humano para Estudos de Mineração de Dados. In *Anais do Congresso Brasileiro de Informática em Saúde*. Santos, SP, Brazil, 2014.

Roddick, J. F. and Spiliopoulou, M. A Survey of Temporal Knowledge Discovery Paradigms and Methods. *IEEE Transactions on Knowledge and Data Engineering* 14 (4): 750–767, 2002.

Rubin, D. B. Conceptual, Computational and Inferential Benefits of the Missing Data Perspective in Applied and Theoretical Statistical Problems. *Allgemeines Statistisches Archiv* 90 (4): 501–513, 2006.

Shahnawaz, M., Ranjan, A., and Danish, M. Temporal Data Mining: An Overview. *International Journal of Engineering and Advanced Technology*, 2011.

Silva, L. O. and Zárate, L. E. A Brief Review of the Main Approaches for Treatment of Missing Data. *Intelligent Data Analysis* 18 (6): 1177–1198, 2014.

United Nations. World Population Prospects: the 2015 Revision, Key Findings and Advance Tables. In *Working Paper ESA/P/WP.241*, 2015.

Williams, G., Baxter, R., He, H., Hawkins, S., and Gu, L. A Comparative Study of RNN for Outlier Detection in Data Mining. In *Proceedings. 2002 IEEE International Conference on Data Mining*. Melbourne, FL, USA, pp. 709–712, 2002.