SN Social Sciences (2024) 4:144
https://doi.org/10.1007/s43545-024-00954-1

ORIGINAL ARTICLE

A hybrid approach to analysing large scale surveys: individual values, opinions and perceptions

Salvatore Flavio Pileggi[1]

Received: 22 April 2024 / Accepted: 24 July 2024 / Published online: 5 August 2024

Abstract

High-dimensional large scale surveys enable broad research capabilities and potential insight. However, when dealing with the intrinsic complexity of social science, the underlying knowledge engineering process may play a critical role and require to consider the characteristics and peculiarities of a given problem in context. This study proposes an analysis framework based on clustering techniques, which have been applied to discover patterns among a number of abstracted features resulting from selected attributes of the World Values Survey (WVS). As an assumption, such features have been softly classified as values, opinions and perceptions, based on their theoretical likelihood to change along the time. From a more philosophical perspective, this work assumes hybrid practices as there is no pre-formulated theory but rather an attempt to discover patterns and new knowledge from data. Given the relatively manageable dimensionality of the input dataset, the feature selection has been performed according to an application-oriented approach (rather than driven by statistical analysis) to establish a more comprehensive and consistent research framework. Among the main findings, a symbiotic relationship between the perception of satisfaction and of financial stability, as well as between the perception of security and of happiness, in addition to more complex patterns involving traditional values (e.g. family and religion), politics and believes with an impact on society. Last but not least, despite its holistic focus, the study has allowed the identification of few research gaps and, therefore, potential further research direction in the broad domain of Social Sciences.

Keywords Human behaviour · Social indicators · Computational social science · Hybrid science · Data analysis · Unsupervised learning · Clustering · Knowledge engineering

✉ Salvatore Flavio Pileggi
SalvatoreFlavio.Pileggi@uts.edu.au

[1] University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

Introduction

The World Values Survey (WVS) is "an international research program devoted to the scientific and academic study of social, political, economic, religious and cultural values of people in the world".[1] In order to sustain such an ambitious goal, a representative comparative social survey (Haerpfer et al. 2022) is conducted globally every 5 years. The survey data is freely available and, by enabling an effective open data philosophy (Murray-Rust 2008), it progressively became one of the most authoritative and widely-used cross-national surveys in the broad field of social sciences.

The enormous popularity and relevance of WVS within the scientific community is evident looking at the enormous number of studies in literature based on such survey data. Just to provide few significant examples among the very many, WVS allows cross-national comparison (Alemán and Woods 2016), agile analysis (e.g. MacIntosh 1998; Silver and Dowley 2000; Bruni and Stanca 2006; Amoranto et al. 2010; Cowley and Smith 2014) and easy extension or deepening (e.g. Johnson and Mislin 2012). Because of the multidimensionality of the dataset, resulting studies may have a more holistic focus (e.g. Fleche et al. 2012), as well as they can be framed within a more specific domain, such as education (Koshy et al. 2023) and religion (Freese 2004).

In this study clustering techniques (Rui and Wunsch 2005) have been applied to discover patterns among a number of selected features from WVS. Clustering analysis has been extensively used in a scientific context and keep evolving as a response to the constantly evolving environment (Wierzchoń and Kłopotek 2018).

WVS deals with a large number of values but doesn't provide a formal multi-perspective classification but rather a structure in thematic sections. Additionally, the cross-cultural focus, which is unquestionably one of its major strengths, should be properly considered addressing holistic studies. A basic classification to distinguish between perceptions and more deep-rooted values can be established, at least in generic terms, according to straightforward and relatively objective criteria, in line with definitions and studies in literature. However, such a simplified approach hides an intrinsic underlying complexity. For instance, some perceptions (e.g. "happiness" or "satisfaction") are relatively easy to identify, while others (e.g. "perception of security") may depend on environmental and contextual factors. It applies also to the different values, whose interpretation and consequent classification can vary considerably depending on the analysis context. In this specific work, as an assumption, features have been softly classified as values, opinions and perceptions, based on their theoretical likelihood to change along the time. Such an assumption enables a semantically enriched conceptual framework that does not affect numerical results but enhance qualitative and critical analysis. Additionally, the focus is on believes and opinions that are more likely to have a concrete social impact, such as potentially discriminatory or divisive.

The proposed approach, which combines clustering techniques with a semantically enriched framework, enables a further level of abstraction for critical analysis. At the same time, it naturally facilitates continuity with possible future work, as well as a natural bridge to predictive models. Indeed, the actual scientific value of the

[1] World Values Survey (WVS) - https://www.worldvaluessurvey.org/ - Accessed: 20 November 2023.

identified patters, depends on the capability of interpretation in context through an enhanced knowledge engineering process.

From a more philosophical perspective, this work can be framed within a hybrid context as there is no pre-formulated theory but rather an attempt to discover patterns and new knowledge from data by adopting hybrid practices (Tuunainen 2005). Given the relatively manageable dimensionality of the input dataset, the feature selection has been performed according to an application-oriented approach (rather than driven by statistical analysis) to establish a more comprehensive and consistent research framework.

The study primarily aims at a self-contained analysis based on the identification of critical patterns. On the other side, the application of unsupervised learning techniques (Alloghani et al. 2020), which by definition work on unlabeled data, provides potential for classification and for a natural evolution towards predictive models. The main findings are briefly discussed by facilitating a concise dialogue with literature. It allows an interpretation in context, as well as the identification of possible gaps.

From a methodological perspective, the aimed holistic analysis has been conducted according to a systematic approach, which is described in detail in the next section both with the main research design decisions. As explained later on in the paper, the most critical and sensitive parameter is the set of thresholds to identify patterns. Indeed, since patterns are defined on relative differences among clusters, thresholds become a determinant. It applies to the analysis itself given the methodology adopted, as well as to support additional Machine Learning steps through automated/semi-automated data labelling (e.g. Willemink et al. 2020) and related further application, such as predictive (e.g. Di Francescomarino et al. 2016, among the very many) and ontological modelling (for example Pileggi 2023a). Instead of considering such a parameter as a variable with a consequent impact on the clarity of the analysis, we have defined thresholds looking at the adopted numerical scale and defined regular numerical steps accordingly. We believe that such an approach foster a more transparent and understandable communication of the key findings.

Looking holistically at this research, the exploratory focus and its inherent simplicity naturally enable insight and added value for future theoretical development. The latter can be translated into enhanced capabilities to bridge existing knowledge gaps, as well as to identify additional gaps and outline research directions accordingly.

Research question(s)

In more formal terms, the paper addresses the following research questions:

- *How to systematically enable hybrid science looking at large datasets?*
- *How can such methods be applied to complex case studies to generate insight and added value?*

Structure of the paper

The core part of the paper is structured according to a classic schema that assumes design and methodological aspects presented in section "Research Design, methodology and approach", computation results briefly summarised in section "Computation results" and discussed in section "Discussion".

Research design, methodology and approach

As previously introduced, the study adopts clustering techniques and has been conducted by following the process depicted in Fig. 1. This section explains such a process in detail with a focus on the research design and key related decisions.

Dataset

This study is based on the seventh wave of the WVS survey (2017–2021) (JD Systems Institute & WVSA 2022), which is composed of 290 core questions (including also demographics) in addition to a number of extra modules. Core questions are structured in 13 different thematic sections, in addition to demographics. The original dataset includes 94,278 answers to the survey.

As a rich data source, WVS makes possible in practice studies with a broad scope and purpose with a different focus, including, among others, cross-cultural or cross-country and long-term trend analysis. At the same time, WVS supports more specific studies as many contributions in literature and previously mentioned works demonstrate.

This study keeps a generic and holistic focus in principle by dealing with multiple values, related dependencies and social implications. However, the analysis is conducted in the context of an abstracted framework which distinguishes between deep rooted values, believes/opinion and perceptions.

Feature selection

In general terms, given the dimensional richness of the original dataset, statistical analysis can potentially provide insight and contribute to identify key patterns. How-

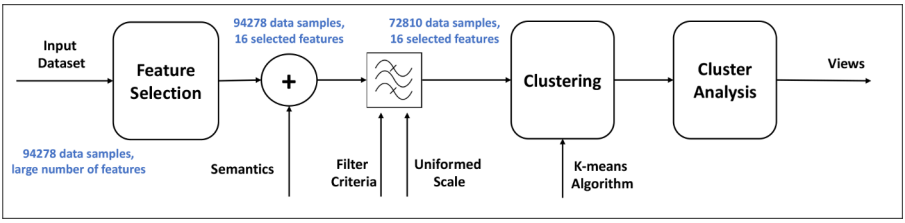

Fig. 1 Overview of the process

ever, rather than on a holistic analysis of the dataset, this work focuses on the analysis of specific features of interest that defines a subset f of the original dataset F (Eq. 1).

$$f = [f_0, f_1, f_2, \ldots, f_i] \qquad f_k \in F \;\; \forall\, k \in [0, \ldots, i] \implies f \subseteq F \tag{1}$$

Feature selection plays a critical, if not determinant, role in many context and studies. In a computational world that wants normally to take advantage of data with very high dimension, systematic approaches are normally requested and, indeed, a number of consolidated algorithms for feature selection have been developed by the research community (Chandrashekar and Sahin 2014). Additionally, certain disciplines and applications may even suggest a more domain-specific approach (e.g. bio-informatics (Saeys et al. 2007)).

In this specific context, the goal is to maximizing the application value rather than the number of identified statistical correlations. Therefore, taking advantage of the manageable number of features, features have been selected as part of a modelling process according to the following criteria:

- Selected features should independently model *stand-alone attributes*. As the original survey is quiet structured and questions are often grouped, this criterion becomes critical and enables relatively clear boundaries.
- Features are selected trying to *minimize the potential overlapping*. For the same reasons previously mentioned, certain questions are addressing similar concepts.
- The priority is on *generalizable attributes*—i.e. those attributes that are more likely to reflect generic concepts at a social level. Indeed, the abstracted conceptualization introduced by the analysis framework needs to be consistent with the underpinning data and semantically consolidated. In line with the first criterion, selected questions should be on aspects of general relevance at a social level, for instance because potentially discriminatory or divisive.

The application of such criteria has led to the selection of 16 features. The original IDs and related survey questions are reported in Table 1. Because of the holistic research focus, no demographic feature has been considered. Demographics could be considered to address more specific studies.

Because of the generality of the selection principles, the feature selection process presents a certain degree of subjectivity on the potential relevance of the different features in the context of the proposed study. This is a relatively common situation, for instance when composite indicators are generated from more specific ones (e.g. in Pileggi 2019, 2022, 2023b). However, in this specific case, we believe that the bias is limited by the scope of the analysis conducted.

Semantic characterization

Each selected feature defines a concept as reported in Table 1. Additionally, as shown in the same table, concepts are classified as follows:

Table 1 Features and semantic characterization

ID*	Survey question*	Concept	Type	Metric/Scale*-**
Q1	*"… indicate how important it is in your life (Family)"*	***Family***	Value/Principle	Importance/1-4
Q2	*"… indicate how important it is in your life (Friends)"*	***Friends***	Value/Principle	Importance/1-4
Q3	*"… indicate how important it is in your life (Leisure)"*	***Leisure***	Value/Principle	Importance/1-4
Q4	*"… indicate how important it is in your life (Politics)"*	***Politics***	Value/Principle	Importance/1-4
Q5	*"… indicate how important it is in your life (Work)"*	***Work***	Value/Principle	Importance/1-4
Q6	*"… indicate how important it is in your life (Religion)"*	***Religion***	Value/Principle	Importance/1-4
Q27	*"One of my main goals in life has been to make my parents proud"*	***Parents Opinion***	Value/Principle	Agreement/1-4
Q29	*"On the whole, men make better political leaders than women do"*	***Gender Discrimination***	Opinion/Belief	Agreement/1-4
Q36	*"Homosexual couples are as good parents as other couples"*	***Homosexuality Acceptance***	Opinion/Belief	Agreement/1-5
Q46	*"Taking all things together, would you say you are (happy)"*	***Happiness***	Perception	Perception/1-4
Q49	*"How satisfied are you with your life as a whole these days?"*	***Satisfaction (overall)***	Perception	Satisfaction/1-10
Q50	*"How satisfied are you with the financial situation of your household?"*	***Financial Stability***	Perception	Satisfaction/1-10
Q60	*"Could you tell me for each whether you trust people from this group…? (People you know personally)"*	***Trusting in others***	Opinion/Belief	Trust/1-4
Q69	*"Could you tell me how much confidence you have in… (Police)"*	***Confidence in Authorities***	Opinion/Belief	Trust/1-4
Q112	*"How would you place your views on corruption in your country"*	***Corruption***	Perception	Perception/1-10
Q131	*"Could you tell me how secure do you feel these days?"*	***Security***	Perception	Perception/1-4

*As in the original dataset (JD Systems Institute & WVSA 2022)

**Lower values correspond to higher levels (e.g. the value 1 indicates a higher "consideration" than the value)

- *Demographic* as in a common meaning.
- *Value*, understood as a principle or standard of behaviour. In this context, the key assumption is that, in very generic terms, values are unlikely to change very much at an individual level as they normally result from the cultural background and other radicated factors.
- *Opinion/Belief*, as in a common meaning. The assumption is that an opinion is still a firm individual believe but it is somehow more likely to change than a value.

- *Perception*, a belief or opinion that significantly differs from the previous category as it is based on a very personal, often temporary, understanding or interpretation of a given reality or situation. It is assumed to be much more volatile than the previous category as a perception can change relatively often in response to happenings or changes of circumstances.

That is understood to be a very soft classification (summarised in Fig. 2) as there is no clear boundary among value, opinion/belief and perceptions. Additionally, such concepts are adopted across different studies with slightly different semantics.

As previously discussed, a completely objective classification is unrealistic. However, a number of driving criteria have been applied in this specific study.

First of all, even in a context of continuous cultural change and evolution (Inglehart and Baker 2000), traditional values are more luckily to preserve their deep-rooted character. It is definitely the case of religion (Roccas 2005; Luckmann et al. 2022), family (Hakim 2018) (often extended to friends (Pahl and Pevalin 2005)) and related opinions (Knafo and Galansky 2008), work culture (Casey 2013) and its balance with personal life (Brough et al. 2020), as well as engagement/interest in politics (Bowler et al. 2007). Such considerations have driven the identification of a corpus of values/principles (*Q1*-*Q6* and *Q27* in Table 1).

On the other side, a set of perceptions has been identified assuming the commonly accepted definition and their potential social relevance (e.g. Oppenheimer 2006). While perceptions may present some dependency on cultural and deep-rooted factors, their subjectivity and potential likelihood to change in response to personal and/or environmental factors define a trend in contrast, if not opposite, with the previous category, mostly characterized by "stability". The features associated with this class in the study (*Q46, Q49, Q50, Q112* and *Q131* in Table 1) are considered as perceptions in a large number of studies in literature. Happiness (e.g. Robert Cloninger and Zohar 2011) and life satisfaction (e.g. Miller et al. 2019) are constantly object of study within the scientific community, as well as financial well-being (e.g. Ponchio et al. 2019), corruption (Melgar et al. 2010) and security in the very many possible meaning (e.g. Greco and Polli 2021). To note that, in general terms, there is a significant difference between a perception and a more or less official measure or estimation. A typical example is corruption (Olken 2009).

The additional category (opinion/belief) presents hybrid characteristics as it includes features that are reasonably related to the cultural background and education but also sensitive of the social environment. The underlying complexity is exten-

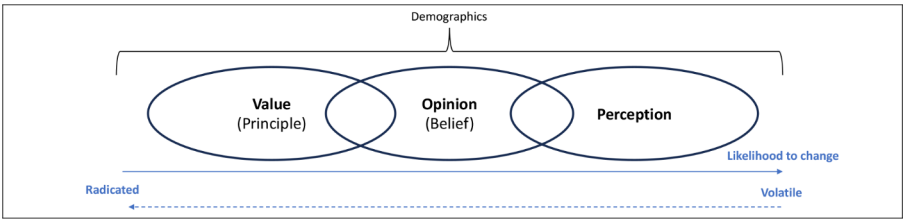

Fig. 2 Soft classification of concepts

sively discussed in literature. It's the case of the different aspects of homosexuality acceptance (e.g. La Roi and Mandemakers 2018), the very many kind of gender discrimination (for instance at work (Cleveland et al. 2013)), social trusting (Holmberg and Rothstein 2017) and confidence in authorities (Tyler 2001). These features have been selected in this study (*Q29, Q36, Q60* and *Q69* in Table 1).

Filtering and scaling

Incomplete data entries are not suitable for the target analysis. Filtering by accepting only positive values allows the inclusion of complete line only as the original dataset codification assumes negative values for missing information. Filtering is formalised in Eq. (2), where X^f represents the set of multi-dimensional data point x, dimensionally restricted to the set of selected features f. A given data point x_j is considered in the study whether its value is positive for each selected feature. After filtering, valid data entries are 72,810.

$$x_j^f = \left[x_j^{f_0}, x_j^{f_1}, x_j^{f_2}, \ldots, x_j^{f_i}\right] \in X^f \qquad f_0, \ldots, f_i \in f \qquad x_j^{f_i} > 0 \qquad \forall i, j \tag{2}$$

Because of the different scales (Table 1), the formula in Eq. (3) is applied to obtain a uniform representation between a minimum value $f^{Min} = 1$ and a maximum value $f^{Max} = 4$. That is probably the most natural choice because most selected features (12 out of 16) are expressed according to that scale in the original dataset as reported in Table 1.

$$x_j^{f_i} = f^{Min} + \frac{(x_j^{f_i} - min(x^{f_i})) * (f^{Max} - f^{Min})}{max(x^{f_i}) - min(x^{f_i})} \qquad [f^{Min}, f^{Max}] = [1, 4] \tag{3}$$

Clustering

Unsupervised Learning (Alloghani et al. 2020) is a branch of Machine Learning that learns from unlabeled data and, therefore, without human supervision. It is often adopted to explore data, discover patterns and new knowledge, as well as to prepare further learning steps. Clustering algorithms (Rui and Wunsch 2005) deal with the data structure partition in unknown area according to different metrics and processes (Xu and Tian 2015).

The different clustering techniques are typically classified depending on their underlying approach. An effective categorization assumes five major classes of solutions (hierarchical, partitional, grid, density-based and model-based) (Saxena et al. 2017). An exhaustive discussion of the different solutions is out of the scope of this paper.

The analysis conducted is based on the classic k-means algorithm for clustering (Sinaga and Yang 2020), which belongs to the partitional clustering category as no hierarchical structure is assumed. K-mean is probably the most popular and bench-

marked algorithm in the field (Saxena et al. 2017). Because of its simplicity that somehow fosters transparency, it is ideal in this specific context dealing with a need for generic clustering without specifically critical requirements. More concretely, the computations adopt the Scikit-learn python package (Pedregosa et al. 2011), which provide an implementation of k-means within an integrated framework. Such a software library is very popular within the scientific community as it is freely available and considered to be highly reliable. Similar criteria applies to determine the optimal number of clusters, that has been estimated heuristically by adopting the popular elbow method (Thorndike 1953) with the support of the Yellowbrick package (Bengfort and Bilbro 2019). By providing a kind of saturation value, i.e. a point where diminishing returns are not worth the corresponding increasing in cost or complexity, the method is simple and intuitive, as well as it is freely available as a computational resource.

Cluster analysis

The analysis is based on centroids, which are understood as the centres of the identified clusters. Centroids are vectors whose values are the mean of each feature. Because of its characteristics, a centroid can be understood as a kind of representative of a cluster.

In order to facilitate the analysis, each centroid is normalised by feature as per Eq. (4), where C^f is the set of centroids and x^k is a single centroid associated with the cluster k. The maximum "interest"/"consideration" is associate with the value 0, while higher values indicates a proportional decreasing of the associated relevance. This simple transformation allows to reason in terms of relative difference.

$$\begin{aligned} x_k^f \in C^f : \quad & x_k = centroid(cluster_k) \\ x_k^{f_i} = x_k^{f_i} - min(x_k^{f_i}) \quad & \forall x_k \in C^f \quad , \ \forall f_i \in f \end{aligned} \tag{4}$$

The matrix resulting by merging the different centroids is provided according to two different views that are numerically equivalent but adopt a different visualization technique:

- the *holistic view* aims to identify overall key patterns. Negligible values (<0.5) are represented in white; values between 0.5 and 1 in yellow; values between 1 and 1.5 in orange and, finally, values higher than 1.5 in red. Such a representation allows to intuitively identify different levels of *consideration or relevance*, including *high* (white), *moderate* (yellow), *low* (orange), *very low* (red). Such qualitative thresholds are summarised in Table 2.

Table 2 Qualitative thresholds for centroid analysis

Numerical value	Qualitative value	Colour
0	*Maximum consideration/relevance*	White
<0.5	*High consideration/relevance*	White
[0.5, 1]	*Moderate consideration/relevance*	Yellow
[1, 1.5]	*Low consideration/relevance*	Orange
>1.5	*Very low consideration/relevance*	Red

- the *feature view* enables a kind of local view to systematically analyse the key patterns related to single features. A gradient scale is applied by feature to put emphasis on the distribution of each feature across the different centroids. Darker colors are associated with higher numerical values and, therefore, with less consideration/importance.

Because of the nature of the proposed views, which express relative patterns, they are considered looking also at major statistics (typically mean and standard deviation) to provide a more consistent interpretation in context. Moreover, qualitative views are always considered in the quantitative context in which they are generated. In other words, the qualitative characterization resulting from a more or less systematic partition of the evaluation space is useful to provide abstraction, while the associated interpretations are still quantitative.

Computation results

Major statistics on the input data are reported in Table 3, which shows the mean, the standard deviation, skewness and kurtosis for each selected feature.

Most features classified as values (Family, Friends, Leisure, Work and Parents Opinion) are highly considered (low mean) in addition to a generalised perception of high happiness (mean = 1.83). Family (mean = 1.12) is the most valued by far. Politics, Religion, Gender Discrimination, Homosexuality Acceptance, Trusting in Others, Confidence in Authorities, Security Perception and financial stability are in a medium range (mean between 2 and 3), with a relatively high standard deviation. On

Table 3 Main statistics on the input data (mean, standard deviation, skewness and kurtosis)

Feature	Mean	Std	Skewness	Kurtosis
Family (Value)	1.12	0.37	3.50	14.26
Friends (Value)	1.70	0.73	0.78	0.13
Leisure (Value)	1.75	0.76	0.75	0.05
Politics (Value)	2.59	0.96	–0.10	–0.94
Work (Value)	1.58	0.79	1.35	1.32
Religion (Value)	2.09	1.11	0.50	–1.18
Parents Opinion (Value)	1.71	0.76	0.83	1.18
Gender Discrimination (Opinion/Believe)	2.72	0.93	–0.33	–0.72
Homosexuality Acceptance (Opinion/Believe)	2.60	0.99	–0.10	–1.17
Trusting in others (Opinion/Believe)	2.06	0.79	0.53	0.05
Confidence in Authorities (Opinion/Believe)	2.39	0.92	0.22	–0.77
Corruption (Perception)	3.20	0.82	–0.80	–0.28
Security (Perception)	2.06	0.81	0.44	–0.26
Happiness (Perception)	1.83	0.69	0.57	0.33
Financial Stability (Perception)	2.77	0.80	–0.41	–0.42
Satisfaction (Perception)	3.06	0.73	–0.73	0.21

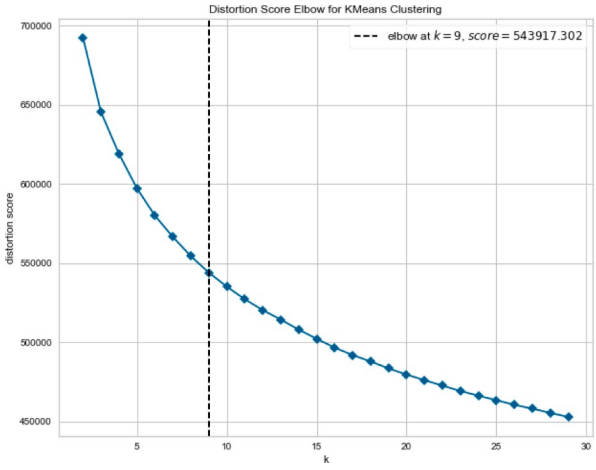

Fig. 3 Heuristic estimation of the optimal number of clusters to consider

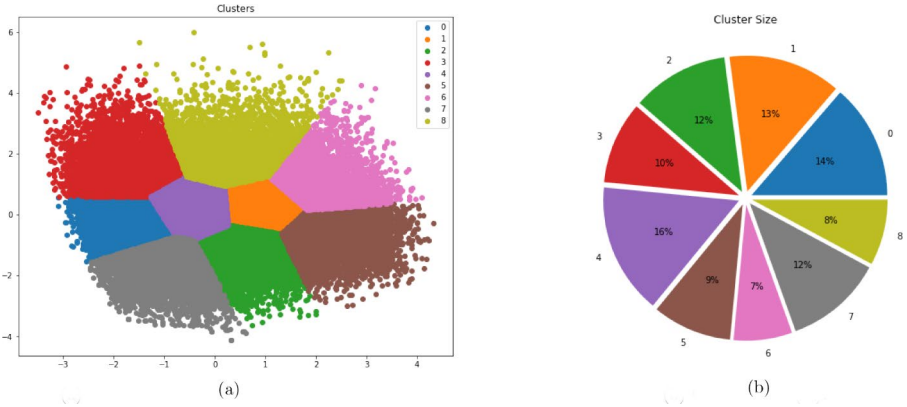

Fig. 4 Visualization of clusters. (**a**) Bi-dimensional visualization of clusters. (**b**) Clusters size (%)

the other extreme, a very low perception of corruption (mean = 3.2) and of satisfaction (mean = 3.06).

In terms of skewness and kurtosis, the input dataset presents positive and negative values in a relatively limited range. The only outstanding exception is family, which has a remarkable right-skewed distribution and high kurtosis value.

The number of clusters to consider is an input for the K-Means algorithm. The optimal number of clusters (k = 9) has been heuristically estimated by applying the *elbow algorithm*. In this specific case, such an optimization point is considered to be reasonable in context, looking at the actual dimensionality—i.e. at the number of considered features. A visualization of the outcome based on the *distortion score* is reported in Fig. 3.

The computation outcome is visualized in Fig. 4: a bi-dimensional representation is proposed in Fig. 4a, while Fig. 4b presents the size of the different clusters in percentage. As shown in the figure, clusters are relatively balanced in size in a range [7–16%].

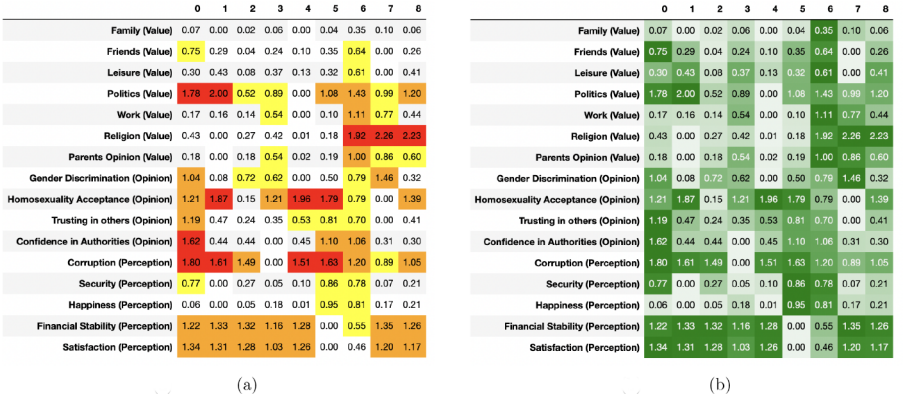

Fig. 5 Holistic and feature view. (**a**) Holistic view. (**b**) Feature view

The *Holistic View* and the *Feature View* as previously defined are reported in Fig. 5a and 5b respectively. They are two dimensional structures, where columns are clusters and row features. Therefore, a column indicates the values of the different features for a given centroid. Key findings are identified on these views and are discussed in detail in the following section.

Discussion

A cluster analysis facilitated by the provided views allows the identification of several patters of potential interest. Such patterns largely depends on relative thresholds, their interpretations, as well as on the specificity of the focus of the study. In this section, major patterns, identified assuming a relatively low number of thresholds (Table 2), are described and discussed in context with an holistic focus. Such findings are summarised in Table 4.

There is an evident symbiotic relationship existing between the perception of satisfaction and of financial stability, as the patterns of the two features (Fig. 5) present a very high degree of similarity. A possible correlation is somehow suggested by the similar mean and standard deviation (Table 3), which averagely point out very low levels of perceived satisfaction and financial stability. The cluster analysis provides further insight and it is largely supported by literature, since the relationship between financial and life satisfaction has been often object of study. For instance(Medgyesi and Zólyomi 2016) addresses the explicit impact of job and financial satisfaction on the overall satisfaction with life, (Christoph 2010) suggests a more accurate analysis by adopting alternative measures, (Gray 2014) adopts a different analysis strategy focusing on financial concerns, (Boes and Winkelmann 2010) puts specific emphasis on the impact of the income, (Diener and Biswas-Diener 2002) is characterised by a more holistic focus, while (Frijters et al. 2004) provides empirical evidences.

A similar relationship exists also between the perception of happiness and of security (Fig. 5a), in this case with absolute values that may be considered in a medium range (Table 3). In general terms, the definition of security is contextual. In the origi-

Table 4 Summary of the major identified patterns by cluster analysis

Short description	Involved feature(s)	Cluster(s)
Symbiotic relationship between the perception of overall satisfaction and of financial stability.	Satisfaction, Financial Stability	[0–8]
Symbiotic relationship between the perception of happiness and of security.	Happiness, Security	[0–8]
Religion presents the most polarised patterns.	Religion	[0–8]
Politics, perception of corruption and homosexuality acceptance are the most regularly distributed features.	Politics, Corruption, Homosexuality Acceptance	[0–8]
A maximum interest in politics (cluster 4) is correlated to a very high perception of corruption and a very low acceptance of homosexuality.	Politics, Homosexuality Acceptance, Corruption	4
A very low level of homosexuality acceptance (cluster 1,4 and 5) is related, among others, to a high consideration of religion, while the higher levels of acceptance (cluster 2 and 7) are associated with both a low (cluster 7) and a high (cluster 2) consideration of religion.	Homosexuality Acceptance, Religion	1, 2, 4, 5, 7
Family is value appreciated the most.	Family, Friends, Leisure	[0–8]
Happiness and satisfaction present different patterns.	Happiness, Satisfaction	[0–8]

nal survey, security is approached holistically, both in the formulation of the question itself and in the context of the corresponding section that includes multiple questions. Also in this case, navigating the literature may result in a very articulated process involving many factors (e.g. Ouweneel 2002), as well as the self-assessed perception of happiness (Dolan et al. 2008) is far way to be uniquely understood and may depend on different determinants (Schimmel 2009).

Additionally, the conducted analysis allows some considerations about the relationship between happiness and satisfaction. Although in certain contexts the two concepts are indistinctly used or, anyway, considered to be similar, they have a different definition as "happiness is a momentary experience that arises spontaneously, while life satisfaction is a long-term feeling based on achieving life-long goals" (Badri et al. 2022). In this specific case, there is averagely a much higher level of perceived happiness (mean = 1.83) than of satisfaction (mean = 3.06). Cluster analysis has pointed out different patterns (Fig. 5). The existing literature demonstrates the complexity of the relationship, which normally requires a multi-domain analysis to be properly addressed (Michalos 1980). For example, the specific role of work has been investigated by the work proposed in (Dockery et al. 2003), as well as a more generic study (Peiró 2006) frames the analysis within the broad socio-economic conditions.

Religion presents by far the most polarised patters with six out of the nine clusters that consider it to be a relevant value and the remaining clusters associating a low relevance (Fig. 5a). In numerical terms, that is consistent with the main statistic reported in Table 3 (mean in a medium range and high standard deviation). This is in a way re-iterating the relevance and the role of religion in a constantly evolving society (Turner 2011; Luckmann et al. 2022).

Politics, perception of corruption and homosexuality acceptance present the opposite trend, as they are the most regularly distributed across the different clusters. That is evident in Fig. 5a since all the qualitative characterizations reported Table 2 are present for these features. The mean associated is high for the three attribute, meaning that averagely there is a low interest in politics, a low level of homosexuality acceptance and a high perception of corruption. Looking at the relationship among these features, the most evident pattern (cluster 4) associates a maximum interest in politics with a very high perception of corruption and a very low acceptance of homosexuality. While the perception corruption is extensively addressed in literature, as far as we know, there is a much more limited knowledge on the relationship between political interest and perception of corruption (e.g. Dong and Torgler 2009). Similarly, we could not identify any specific study on the relationship between interest in politics and homosexuality acceptance.

Focusing more specifically on homosexuality acceptance, a very low level of homosexuality acceptance (cluster 1,4 and 5) is related, among others, to a high consideration of religion, while the higher levels of acceptance (cluster 2 and 7) are associated with both a low (cluster 7) and a high (cluster 2) consideration of religion. Such a relationship is expensively addressed in literature (Adamczyk and Pitt 2009; Whitehead and Baker 2012; Jäckle and Wenzelburger 2015; Xie and Peng 2018).

Finally, family is value appreciated the most. Also other values (work, friends, leisure and parents opinion) are generally appreciated, although at a slightly minor level.

Conclusions and future work

The application of unsupervised learning techniques and enriched semantics on a large scale survey has enabled a dynamic analysis framework according to an hybrid science approach aimed at discovering patters among features of interest. The analysis of the resulting clusters has provided insight and, more in general, added value. A number of critical patterns have been identified accordingly and discussed by facilitating a dialogue with literature. Because of the high-dimensionality of the input dataset reflecting a variety of related social aspects, a clear identification of most critical patters may be challenging, especially in a cross-cultural context. The methodology adopted and the associated framework on one side provide a computational resource to enable a systematic analysis; on another side allows customisation through semantics to be re-use in a different context.

Furthermore, the hybrid approach characterising this study has contributed to define and progressively consolidate few research gaps and possible future research directions in the field of social sciences, as not all identified patterns seem to be

extensively addressed in literature. Overall, the exploratory focus has resulted to be effective in the context of a complex case study by leveraging on the flexibility of clustering techniques, as well as on the underlying conceptualisation to support future theoretical development.

The major limitation of the analysis conducted is related to a fundamental sensitivity of the adopted thresholds, that have been intuitively defined looking at numerical scales. As previously discussed, other potentially biasing factors in the research design can be considered in a context of theoretical trade-off to establish a hybrid approach in fact. In the specific use case addressed in this study, human (feature selection) and computational-driven (number of clusters) design decisions converge towards an heuristic optimization of such a trade-off, which assumes, ideally, low dimensionality and a proportional number of clusters. The dimensionality reduction performed by conceptualization finally resulted in 16 selected features associated with 9 clusters. Therefore, the ratio features/clusters is close to 2. It can be intuitively considered to be a good configuration for the proposed analysis framework at a low scale.

The holistic focus of the analysis intrinsically suggests more specific studies, as well as a consolidation of the main findings, for instance considering a multi-feature approach for the conceptualization. Additional statistical test to confirm the strengths of the identified relationships should be conducted on more extended data spaces resulting from the integration of different datasets. The most natural direction for future work is probably the generation of predictive models based on the provided classification that allows data labelling and the consequent learning loop closing.

Acknowledgements I would like to thank the team behind the World Value Survey for making the dataset freely available for the community. Additionally, a sincere acknowledgement to the anonymous reviewers, who have provided valuable feedback.

Author contributions The paper has one single author.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research was not externally funded.
Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability This work is based exclusively on secondary data. The original dataset (JD Systems Institute & WVSA 2022) is cited in the paper and is freely available.

Declarations

Ethical approval Not applicable.

Informed consent Not applicable.

Conflict of interest The author declares that he has no competing interests.

References

Adamczyk A, Pitt C (2009) Shaping attitudes about homosexuality: the role of religion and cultural context. Soc Sci Res 38(2):338–351

Alemán J, Woods D (2016) Value orientations from the world values survey: how comparable are they cross-nationally? Comp Polit Studi 49(8):1039–1067

Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ (2020) A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Supervised and Unsupervised Learning for Data Science, pp 3–21

Amoranto G, Chun N, Deolalikar AB (2010) Who are the middle class and what values do they hold? Evidence from the world values survey. Evidence from the World Values Survey (October 1, 2010) Asian Development Bank Economics Working Paper Series 229

Badri MA, Alkhaili M, Aldhaheri H, Yang G, Albahar M, Alrashdi A (2022) Exploring the reciprocal relationships between happiness and life satisfaction of working adults—evidence from Abu Dhabi. Int J Environ Res Public Health 19(6):3575

Bengfort B, Bilbro R (2019) Yellowbrick: visualizing the Scikit-learn model selection process. J Open Source Softw 4(35):1075

Boes S, Winkelmann R (2010) The effect of income on general life satisfaction and dissatisfaction. Soc Indic Res 95:111–128

Bowler S, Donovan T, Karp JA (2007) Enraged or engaged? preferences for direct citizen participation in affluent democracies. Political Res Q 60(3):351–362

Brough P, Timms C, Chan XW, Hawkes A, Rasmussen L (2020) Work–life balance: definitions, causes, and consequences. In: Handbook of socioeconomic determinants of occupational health: from macro-level to micro-level evidence, pp 473–487

Bruni L, Stanca L (2006) Income aspirations, television and happiness: evidence from the world values survey. Kyklos 59(2):209–225

Casey C (2013) Work, self and society: after industrialism. Routledge

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Christoph B (2010) The relation between life satisfaction and the material situation: a re-evaluation using alternative measures. Soc Indic Res 98:475–499

Cleveland JN, Vescio TK, Barnes-Farrell JL (2013) Gender discrimination in organizations. In: Discrimination at work. Psychology Press, pp 177–204

Cowley E, Smith S (2014) Motivation and mission in the public sector: evidence from the world values survey. Theory Decis 76:241–263

Di Francescomarino C, Dumas M, Maggi FM, Teinemaa I (2016) Clustering-based predictive process monitoring. IEEE Trans Serv Comput 12(6):896–909

Diener E, Biswas-Diener R (2002) Will money increase subjective well-being? Soc Indic Res 57:119–169

Dockery AM et al (2003) Happiness, life satisfaction and the role of work: evidence from two Australian surveys. School of Economics and Finance, Curtin University of Technology

Dolan P, Peasgood T, White M (2008) Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. J Econ Psychol 29(1):94–122

Dong B, Torgler B (2009) Corruption and political interest: empirical evidence at the micro level. J Interdiscip Econ 21(3):295–325

Fleche S, Smith C, Sorsa P (2012) Exploring determinants of subjective wellbeing in oecd countries: evidence from the world value survey

Freese J (2004) Risk preferences and gender differences in religiousness: evidence from the world values survey. Rev Relig Res 88–91

Frijters P, Haisken-denew JP, Shields MA (2004) Money does matter! evidence from increasing real income and life satisfaction in east germany following reunification. Am Econ Rev 94(3):730–740

Gray D (2014) Financial concerns and overall life satisfaction: a joint modelling approach

Greco F, Polli A (2021) Security perception and people well-being. Soc Indic Res 153(2):741–758
Haerpfer C, Inglehart R, Moreno A, Welzel C, Kizilova K, Diez-Medrano J, Lagos M, Norris P, Ponarin E, Puranen B (eds) (2022) World values survey: round seven - country-pooled datafile version 5.0. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria. https://doi.org/10.14281/18241.20
Hakim C (2018) Models of the family in modern societies: ideals and realities. Routledge
Holmberg S, Rothstein B (2017) Trusting other people. J Public Affairs 17(1–2):e1645
Inglehart R, Baker WE (2000) Modernization, cultural change, and the persistence of traditional values. Am Sociol Rev 65(1):19–51
Jäckle S, Wenzelburger G (2015) Religion, religiosity, and the attitudes toward homosexuality—a multilevel analysis of 79 countries. J Homosex 62(2):207–241
JD Systems Institute & WVSA (2022) European Values Study and World Values Survey: joint EVS/WVS 2017–2022 Dataset (Joint EVS/WVS). Dataset Version 4.0.0. https://doi.org/10.14281/18241.21.
Johnson ND, Mislin A (2012) How much should we trust the world values survey trust question? Econ Lett 116(2):210–212
Knafo A, Galansky N (2008) The influence of children on their parents' values. Soc Personal Psychol Compass 2(3):1143–1161
Koshy P, Cabalu H, Valencia V (2023) Higher education and the importance of values: evidence from the world values survey. Higher Educ 85(6):1401–1426
La Roi C, Mandemakers JJ (2018) Acceptance of homosexuality through education? investigating the role of education, family background and individual characteristics in the United Kingdom. Social Sci Res 71:109–128
Luckmann T, Kaden T, Schnettler B (2022) The invisible religion: the problem of religion in modern society. Routledge
MacIntosh R (1998) Global attitude measurement: an assessment of the world values survey postmaterialism scale. Am Sociol Rev 452–464
Medgyesi M, Zólyomi E (2016) Job satisfaction and satisfaction in financial situation and their impact on life satisfaction, vol 6. European Commission, Directorate general for employment, social affairs and inclusion, pp 2016
Melgar N, Rossi M, Smith TW (2010) The perception of corruption. Int J Public Opin Res 22(1):120–131
Michalos AC (1980) Satisfaction and happiness. Soc Indic Res 8:385–422
Miller BK, Zivnuska S, Michele Kacmar K (2019) Self-perception and life satisfaction. Pers Individ Dif 139:321–325
Murray-Rust P (2008) Open data in science. Nat Precedings 1–1
Olken BA (2009) Corruption perceptions vs. corruption reality. J Public Econ 93(7–8):950–964
Oppenheimer L (2006) The belief in a just world and subjective perceptions of society: a developmental perspective. J Adolesc 29(4):655–669
Ouweneel P (2002) Social security and well-being of the unemployed in 42 nations. J Happiness Stud 3:167–192
Pahl R, Pevalin DJ (2005) Between family and friends: a longitudinal study of friendship choice. Brit J Sociol 56(3):433–450
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Peiró A (2006) Happiness, satisfaction and socio-economic conditions: some international evidence. J Socio Econ 35(2):348–365
Pileggi SF (2019) Is the world becoming a better or a worse place? A data-driven analysis. Sustainability 12(1):88
Pileggi SF (2022) Holistic resilience index: measuring the expected country resilience to pandemic. Qual Quant
Pileggi SF (2023a) Ontological modelling and social networks: from expert validation to consolidated domains. In: International Conference on Computational Science. Springer, pp 672–687
Pileggi SF (2023b) Walking together indicator (wti): understanding and measuring world inequality. Sustainability 15(6):5392
Ponchio MC, Cordeiro RA, Gonçalves VN (2019) Personal factors as antecedents of perceived financial well-being: evidence from Brazil. Int J Bank Mark 37(4):1004–1024
Robert Cloninger C, Zohar AH (2011) Personality and the perception of health and happiness. J Affect Disorder 128(1-2):24–32
Roccas S (2005) Religion and value systems. J Soc Issues 61(4):747–759

Rui X, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517
Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, Er MJ, Ding W, Lin C-T (2017) A review of clustering techniques and developments. Neurocomputing 267:664–681
Schimmel J (2009) Development as happiness: the subjective perception of happiness and UNDP's analysis of poverty, wealth and development. J Happiness Stud 10(1):93–111
Silver BD, Dowley KM (2000) Measuring political culture in multiethnic societies: reaggregating the world values survey. Comp Political Studi 33(4):517–550
Sinaga KP, Yang M-S (2020) Unsupervised k-means clustering algorithm. IEEE Access 8:80716–80727
Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276
Turner BS (2011) Religion and modern society: citizenship, secularisation and the state. Cambridge University Press
Tuunainen J (2005) Hybrid practices? Contributions to the debate on the mutation of science and university. Higher Educ 50:275–298
Tyler TR (2001) Public trust and confidence in legal authorities: what do majority and minority group members want from the law and legal institutions? Behav Sci Law 19(2):215–235
Whitehead AL, Baker JO (2012) Homosexuality, religion, and science: moral authority and the persistence of negative attitudes. Sociol Inq 82(4):487–509
Wierzchoń ST, Kłopotek MA (2018) Modern algorithms of cluster analysis, vol 34. Springer
Willemink MJ, Koszek WA, Hardell C, Jie W, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP (2020) Preparing medical imaging data for machine learning. Radiology 295(1):4–15
Xie Y, Peng M (2018) Attitudes toward homosexuality in china: exploring the effects of religion, modernizing factors, and traditional culture. J Homosex 65(13):1758–1787
Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2:165–193