A new formalism to express and operate on diversity measures of qualitative variables, built in a Hilbert space, is presented. The abstract character of the Hilbert space naturally incorporates the equivalence between qualitative variables and is utilized here to (i) represent the binary character of answers to categories and (ii) introduce a new criterium for choosing between different measures of diversity, namely, robustness against uncertainty. The full potential of the formulation on a Hilbert space comes to play when addressing the reduction of categories problem, a common problem in data analysis. The present formalism solves the problem by incorporating strategies inspired by mathematical and physical statistics, specifically, it makes use of the concept of partial trace. In solving this problem, it is shown that properly normalizing the diversity measures is instrumental to provide a sensible interpretation of the results when the reduction of categories is performed. Finally, the approach presented here also allows for straightforwardly measuring diversity and performing category reduction in situations when simultaneous categories could be chosen.

I. INTRODUCTION

In a wide range of research fields, such as biology [1, 2], economics [3], political science [4], marketing, communications [5, 6] and other social science, the statistical analysis of qualitative variables is fundamental to construct models that predict the behaviour of systems of interest [see, e.g., [3, 6–8]]. There are two kinds of qualitative variables, nominal and ordinal and from the statistical point of view, they should be manipulated differently [9]. Particularly, measuring variability among a set of qualitative variables is different when the set is composed of ordinal or nominal variables [10].

Computing the variability of a set composed of quantitative variables is straightforward: compute, e.g., the standard deviation of the data to know how close or spread the data is distributed [11]. Conversely, for a set of qualitative variables, this is not a trivial matter and it is not clear how the spreading of a given dataset should be formally calculated. In the literature, there can be found a copious number of approaches aimed to calculate the variability of qualitative data [1, 4, 5, 12–18] proposing a different kind of indices to do so. In the last years, some efforts have also been devoted to that direction, e.g., based on the underlying concept of diversity, some of the mathematical expressions used to measure variability have been discussed and justified (cf. e.g., [6, 7, 19]). Some of the efforts also aimed at classifying variability indices based on the structure of the expressions [6, 7] or their type of measurement [19], such as separation variety and disparity. But none of them has been formulated in a basis independent manner.

Two problems related to the sensibility of the measures are addressed here. The first is the robustness of the measures under the presence of noise or uncertainty that may arise, e.g., when analyzing data from a selected fraction of a given community (sample). Hence, having a classification of the variability indices based on their robustness is pertinent when information is incomplete.

The second problem is the influence of the reduction of categories on the variability measurements. Some work in this direction has been done in the past [4, 6]. Specifically, the proposal made in Ref. [4] consists in replacing by zero the proportion of the category that will be reduced and then redistribute the proportions equally among the other categories. From an information science perspective, this strategy violates Landauer’s principle or equivalently, it violates the second law of thermodynamics (see below). Alternatively, in Ref. [6] an approach based on performing a linear regression of the data was suggested, it allows for analyzing how sensible the measures are to the number of categories $k$ and the maximum proportion among categories $P_{\text{max}}$. The Hilbert space formalism presented here leads to a straightforward solution to this problem in terms of the partial trace method [20], that contrary to the previous ones [4, 6], it does not disregard information because when the partial trace is applied the dimension of the resulting state is smaller than the original but keeps all the information provided by the tracedout states.

The dimension of the proportion matrix is increased when mapped onto the Hilbert space, (see below), so that this representation enables the analysis of the variability when several categories may be selected simultaneously, e.g., in a poll when several answers can be chosen at the same time.

II. VARIABILITY MEASURES AND THE DENSITY OPERATOR

Variability is a measure of how spread or localized a set of variables is, i.e., it distinguishes if the frequency of the variables are mainly localized in one variable or if is distributed among them. The dataset can be composed of qualitative or quantitative variables, here, interest is on the former kind. Four different expressions, $(\sigma_L, \sigma_E, \sigma_S, \sigma_P)$, frequently found in the literature [3, 4, 14] are reformulated here to calculate the variability of qualitative
data sets and to allow comparisons among them.

A fundamental aspect towards the formulation of variability measures is the concept of a complete basis set [21, 22]. To be concrete, consider a two-dimensional linear vector space as in Fig. 1. Being elements of a linear vector space, vectors \( \{ \vec{r}_1, \vec{r}_2, \ldots, \vec{r}_{k-1}, \vec{r}_k \} \) are all equivalent and there is no preferable sorting until a particular complete set of basis vectors \( \{ \vec{e}_1, \vec{e}_2 \} \), with unit dyaic \( \vec{1} = \vec{e}_1 \vec{e}_1^\dagger + \vec{e}_2 \vec{e}_2^\dagger \), is selected to represent the initial vector set. The unit vectors \( \vec{e}_j \) is the basis of the dual space [21, 22]. This situation is certainly analog to the situation of qualitative variables. Once the basis is selected, vectors can be sorted, e.g., depending on their projection to any of the basis elements, \( \vec{r}_j \cdot \vec{e}_j \). As any basis set can be utilized, operations over the elements of the space should be independent of the basis. A metric formulation in terms of a basis set with operations that are independent of the basis is the route followed here for qualitative variables.

In doing so, define \( N \) as the number of categories, \( n_i \) the number of answers associated to the \( i \)th -category, \( N \) the total number of answers, i.e., \( N = \sum n_i \) and \( p_i = n_i/N \) the proportion of answers. The central objects towards a Hilbert space formulation are the “density operator" \( \hat{\rho} \) and the complete basis \( \{ |0_1 \ldots 1_j \ldots 0_k \rangle \} \) in the Hilbert space \( \mathcal{H} \). Specifically, for \( k \) categories, the basis element \( |0_1 \ldots 1_i \ldots 0_k \rangle \) represents the situation when the \( i \)th-category was selected with \( p_i = 1 \). Similarly, \( |0_1 \ldots 0_j \ldots 0_k \rangle \) represents the situation when the \( j \)th-category was selected with \( p_j = 1 \). Note that any other complete basis set could be selected; however, the basis set \( \{ |0_1 \ldots 1_i \ldots 0_k \rangle \} \) intrinsically accommodates the proportion of answers to each category. The identity operator reads \( \mathbb{1} = \sum_{i} |0_1 \ldots 1_i \ldots 0_0 \rangle \langle 0_1 \ldots 1_i \ldots 0_0 | \), where the elements \( |0_1 \ldots 1_i \ldots 0_k \rangle \) are elements of the dual of the Hilbert space [21, 22].

The matrix representation of the density operator \( \rho = 1 \cdot \hat{\rho} \cdot 1 \), in the present formulation, corresponds to a diagonal matrix with entries \( p_i \) and with \( \sum_i^{N} p_i = 1 \) (see below). Projections of the density operator onto a particular basis allows for manipulating qualitative and quantitative on the same ground; thus unifying operational tools. The present formulation may be utilized to define probability distributions for qualitative variables in the same form as probability distributions are defined for finite dimensional Hilbert spaces [23, 24]. This may be of great relevance in a wide range of scientific areas such as, political science, marketing, sociology, biology and economics, where the presence of qualitative variables is of great importance. For the later, it is also important a simultaneous manipulation of both, qualitative and quantitative variables to enhance predictability models [25].

In general, the matrix representation of \( \hat{\rho} \) onto the states defined above considers elements of the type \( |0_1 \ldots 1_i \ldots 0_k \rangle \) or \( |1_1 \ldots 1_i \ldots 1_j \ldots 1_k \rangle \) that account for situations when several categories can be selected simultaneously, i.e., situations when categories \( i^\text{th} \) and \( j^\text{th} \) can simultaneously be selected (i.e., element \( |0_1 \ldots 1_i \ldots 0_k \rangle \) or when all the possible categories may be selected (i.e., element \( |1_1 \ldots 1_i \ldots 1_j \ldots 1_k \rangle \)). For simplicity and concreteness, those situations are disregard here by assuming that their corresponding proportion is zero. The resulting elements form a complete set for a subspace that by analogy to quantum mechanics, it will be referred to as the singly-excited manifold and its basis elements as singly-excited states. After projecting the density operator \( \hat{\rho} \), it takes the following form

\[
\rho = \begin{pmatrix}
|0_0_0_0\rangle & |0_1_0_0\rangle & |0_0_1_0\rangle & \cdots & |1_1_1_1\rangle
|0_0_0_0\rangle & |0_1_0_0\rangle & \cdots & |1_1_1_1\rangle
|0_0_0_0\rangle & \cdots & |1_1_1_1\rangle
|0_0_0_0\rangle & \cdots & |1_1_1_1\rangle
|0_0_0_0\rangle & \cdots & |1_1_1_1\rangle
\end{pmatrix}.
\]
Under this consideration, ρ is a diagonal matrix and only one sector is non-zero, namely, the one for which only one category may be selected.

Note that the present formulation resembles the One Hot Encoding Algorithm [26], widely used in Artificial Intelligence literature. However, as it is shown below, the construction based on Hilbert spaces allows for making use, in social sciences, of all analysis tools form, e.g., quantum mechanics. This fact opens the door for a more formal and precise analysis of qualitative variables.

A. Variability measures in terms of the density operator

Once the density operator is projected onto a particular basis, the next step is to calculate its variability in a way that is independent of the representation. This is achieved below by introducing the trace operation, a basis invariant operation, over the density operator ρ. In the literature, it can be found several types of variability measures and attempts to classify them [6, 7, 19]. In this work we are going to focus in the following four:

Measure Type I– The first measure σL is associated with the linear difference between all the answers and can be written as

\[ σ_L = 1 - \frac{1}{2(k-1)} \sum_{\alpha=1}^{k} \sum_{\beta>\alpha}^{k} \text{tr} \left( \hat{ρ} - \Pi_{\alpha\beta} \hat{ρ} \Pi_{\alpha\beta}^\dagger \right), \]  

(2)

where the matrix \( \Pi_{\alpha\beta} \) represents the permutation matrix associated with the rows \( \alpha \) and \( \beta \), i.e., the matrix obtained by permuting the rows \( \alpha \) and \( \beta \) from the identity matrix \( I \) of order \( 2^k \) [27]. The trace operation \( \text{tr}(\cdot) \) represents the sum of the eigenvalues of the matrix and \( \dagger \) refers to the complex transposition.

With a different normalization and without the Hilbert space formulation, an equivalent expression of the Eq. (2) where mentioned before by Wilcoxon [4] as MDA, emphasizing that the main characteristic is that it is “dependent on the spread of the variate-values among themselves and not on the deviations from some central value”.

Measure Type II– The second measurement is similar to the later, but the distance is calculated as an Euclidean distance instead and has the form

\[ σ_E = 1 - \frac{1}{2(k-1)} \left( \sum_{\alpha=1}^{k} \sum_{\beta>\alpha}^{k} \text{tr} \left( \hat{ρ} - \Pi_{\alpha\beta} \hat{ρ} \Pi_{\alpha\beta}^\dagger \right) \right)^\frac{1}{2}. \]  

(3)

An equivalent measure was proposed by Tsui et al [28] in 1992 and named later in 2007 as MED (Mean Euclidean Distance) by Harrison and Klein [19]. This measure also has the same property of the Mean linear distance mentioned by Wilcoxon.

The following two variability expressions are frequently found and used in the qualitative statistics literature, even though it has their ground in the physics literature.

Measure Type III–The Shannon or von Newman entropy [29] can be directly extended to the present formalism after introducing a proper normalization factor. Traditionally, the normalization factor is taken as \( 1/\log_2 k \) that corresponds to the maximum entropy encoded in a density operator of dimension \( k \times k \). Because \( \hat{ρ} \) may include multi-selection of categories, the maximum entropy measurement corresponds to replacing \( k \) in favor of \( 2^k \). Thus,

\[ σ_S = \frac{\text{tr}(\hat{ρ} \log_2 \hat{ρ})}{\text{tr}\left(ρ_{\text{max}} \log_2 ρ_{\text{max}}\right)}, \]  

(4)

where \( ρ_{\text{max}} \) is the density matrix associated to the configuration of maximum entropy of the singly-excited manifold, i.e., a density matrix with the entries on the single excited states equal to \( 1/k \) and the rest of them equal to zero.

In social contexts, the above index was first introduced in the context of behavioral science by Senders [30] as a measure of uncertainty, “which will be high when the number of alternative possibilities is high, and low when some of the possibilities are much more likely than others”.

Measure Type VI– The last measure type, also known as the Index of Qualitative Variation (IQV), is the most commonly used in literature ranging from psychology, politics, economy and others. Its equivalent in physics, particularly in quantum mechanics, is known as linear entropy and corresponds to one minus the purity of the density operator \( \hat{ρ} \), namely, \( 1 - \text{tr}\hat{ρ}^2 \). Specifically, \( σ_P \) takes the following form

\[ σ_P = \frac{1 - \text{tr}\hat{ρ}^2}{1 - \text{tr}\hat{ρ}_{\text{max}}^2}, \]  

(5)

where \( \hat{ρ}_{\text{max}} \) is defined as in Eq. (4). Traditionally, the normalization used for the IQV is \( k/(k-1) \), which also corresponds to the maximum linear entropy of a density matrix with dimension \( k \times k \). In practical terms, the denominators of Eq. (4) and (5) are equal to the traditional normalization factors \(-1/\log_2 k \) and \( k/(k-1) \), respectively, but when the reduction of categories is applied, the selection of a proper normalization factor becomes not trivial (see below).

All previous measures of variability are normalized to 1 so that the limits bounds as 0 and 1. The lower limit is reached when all the categories are zero except one (see Fig. 2) and can be interpreted as no variability at all, and in terms of information, a maximum knowledge of the system. On the other hand, the upper limit appears when all the categories have the same value \( 1/k \), i.e., all of them are equally distributed. Having all the indices bounded by the same values, \([0, 1]\), allows to make direct comparisons among them and also permits a better interpretation of the results. All the indices presented in Eqs. (2) to (5), in some sense, measures how the bars in Fig. 2 are distributed among the categories and also take into account the size of the bars. Interpreting the extremes values of \( σ = 0 \) and \( σ = 1 \) as completely opposite characteristics: Uncompetitive-Competitive, Homogeneous-
TABLE 1. The number of problem-solving courts for the states of Arkansas (Ark.) and Wisconsin (Wis.)

|            | Drug | Mental health | Family | Youth specialty | Hybrid DWI/drug | DWI | Domestic violence | Veterans | Tribal wellness | Other |
|------------|------|---------------|--------|----------------|-----------------|-----|-------------------|----------|----------------|-------|
| Ark.       | 49   | 1             | 0      | 3              | 0               | 0   | 1                 | 1        | 0              | 0     |
| Wis.       | 18   | 3             | 2      | 7              | 8               | 11  | 1                 | 10       | 2              | 1     |

FIG. 2. Pictorial representation of the extreme values of the variability measurement $\sigma$ in a five categories scenario, named \(\{A,B,C,D,E\}\) for one hundred answers. **Left Panel** \(\sigma = 0\), there is no variability among the categories, which means all the information localized in one category. **Central Panel** \(0 < \sigma < 1\), a situation in between is also presented, where the information is “randomly” spread among the categories. **Right Panel** \(\sigma = 1\), the variability is maximum, i.e., the information is maximally spread and equally distributed.

Heterogeneous, Agreement-Disagreement, Segregated-Integrated and Localized-Delocalized, mentioning some of them suggested by Wilcox [4], the variability can be understood as a measure of how equal or unequal is a set of categories.

### B. Robustness of variability measures under uncertainty

All previous measure types provide an idea about how spread the categories are and despite they have the same bounds, their intermediate values may differ. Therefore, interest here is in providing a new classification criterion for selecting, when conceptually possible [6], one measure over the others; specifically, the classification proposal is based on the robustness against uncertainty. Motivation for proposing this criterion is clear, when dealing with real datasets uncertainty is always present, e.g., (i) elasticities of commodity prices with respect to supply or demand [31] represents uncertainty; (ii) the number of violent acts in a society has uncertainty due to the unregistered acts; (iii) when results of a poll are obtained from a sample community there is uncertainty.

In this research, uncertainty was artificially introduced by means of a stochastic variable, i.e., by adding noise to the proportion \(p_i\). Specifically, this is done by adding up a random number \(\zeta_i\) selected from the noise domain. That is to say, e.g., for a noise of amplitude 5%, a random number \(\zeta_1\) is selected between \([-0.05, 0.05]\) and added to the proportion of the first category, \(p_1\); then, another random number \(\zeta_2\) is selected from the same range and added up to the second category, \(p_2\); the same process is repeated until the \((k-1)\)th-category is reached. For the last category, \(p_k\), and with aim of preserve the normalization, \(\sum_{i=1}^{k} p_i = 1\), the number obtained by the summation of the all previous proportions is subtracted from 1. Each set of random numbers \(\{\zeta_1, \zeta_2, \ldots, \zeta_{k-1}\}\) form a realization of the random variable \(\zeta\). Since each number is random, each realization \(\zeta\) is random as well. Therefore, to guarantee reproducibility and to simulate a more realistic situation, thousands of realizations are averaged until no changes above \(10^{-4}\) are detected in the measures, i.e., until convergence is reached. The number of realization to achieve convergence varies from situation to situation and heavily depends on the noise amplitude. For the cases considered below, convergence was reached after averaging out over \(10^{4}\) realizations.

### III. APPLICATION TO CENSUS OF PROBLEM-SOLVING COURTS, 2012

As an application of the formalism introduced here, consider the recently released dataset of the Problem-solving Courts, by State and Selected U.S. Territories, 2012 published by the Bureau of Justice Statistics [32]. The problems solved by the courts in the U.S. are classified in ten different categories, namely, Drug, Mental health, Family, Youth specialty, Hybrid DWI/drug, DWI, Domestic violence, Veterans, Tribal wellness and Other [32]. For the following analysis, each problem is considered as a category with a certain proportion \(p_i\) and variability analysis is performed by each state.

Table I presents two opposite states in terms of the spreading of the problems-solving courts, Arkansas (Ark.) and Wisconsin (Wis.). It can be seen that problems in Arkansas are mainly due to drugs and the other categories have not significant contribution; on the other hand, and even though the main problem in Wisconsin is also drugs, the other categories also have significant values. Hence, Arkansas is expected to have a small variability and contrary, Wisconsin should have high variability values.
The variability values of the four measurements $\sigma_L$, $\sigma_E$, $\sigma_S$ and $\sigma_P$ for the states of Arkansas (Ark.) and Wisconsin (Wis.) are presented in Table II. As anticipated, all measures have a smaller value for Arkansas than for Wisconsin, but the values of the measurements significantly differ among them. Therefore, a comparison can be made among states with the same measurement, but not between measures within the same state.

Figure 3 depicts the values of the four different measures of variability considered above for each territory in the dataset. The horizontal organization of the territories is presented in ascending order for the value obtained by $\sigma_L$. An interesting fact can be addressed here: Although, globally, all variability measures present a growing tendency, locally, behaviour discrepancies are clearly visible. This can be seen comparing the smoothness of the purple line against the oscillation of the others. This implies that a direct comparison of the results among variability measures is not the best way to compare their performance. This is another motivation to highlight that an analysis of robustness of variability measures under uncertainty is a better criterion to select the most accurate and meaningful measure of variability.

Figure 4 depicts the results of the four different measures of variability considered in Fig. 3 with the intensity of the noise 1% (upper panel) and 5% (lower panel) for each of the territories. For the sake of comparison, the continuous curves, on the left-hand-side panel, depict the variability measure for U.S. states without noise whereas the results in the presence of uncertainty are depicted by single markers in the figures (see caption of the figure for more details). The right-hand panel shows the absolute value of the difference between the variability with and without noise for each value of noise intensity. It can be easily seen that the green line is higher than the other measures; therefore, $\sigma_S$ can be interpreted as the most sensitive measure of noise, followed by $\sigma_L$ (purple line). From the picture is not easy to identify whether $\sigma_E$ (blue) or $\sigma_P$ (red) is more sensitive to the noise. Therefore, to quantitative measure the difference between the curves with and without noise, an measure equivalent to the standard deviation is considered for this case. Specifically, assume that the mean value will be given by the no-noise case whereas the noisy situations are to be understood as “experimental” data. The robustness against noise is then quantified by

$$\Phi_x(\theta) = \sqrt{\sum_{n=1}^{\Lambda} (\sigma^n_x(\theta) - \sigma^x_{\theta})^2},$$

(6)

where $\theta$ represent the noise strength (e.g., 1% or 5%), $\Lambda$ is the total number of territories and $x = \{L, E, S, P\}$ labels one of the four possibilities of variability measure.

The values of $\Phi_x(\theta)$ can be found in Table III. The values obtained for $\Phi_E$ are smaller than the others, so it can be concluded that $\sigma_E$ is the most robust measure of qualitative variation.

### IV. VARIATION OF CATEGORIES ANALYSIS

Traditionally, the way to analyze how the reduction of categories affects the variability is by means of the Wilcox’s proposal [4], namely, by replacing by zero the proportion of the reduced category and then renormalize the proportion of the remaining categories. This proposal disregards the information of the reduced categories and assumes that the categories-to-be-reduced do not exist. This formulation violates the basic postulates of information science as well as the second law of thermodynamics because entropy-like measures decrease when information decreases (randomness increases). Below, it is described how the Hilbert space formulation allows performing a reduction of categories while keeping the information provided by the reduced categories and consistent with information science. For simplicity, the reduction of categories analysis will be performed in the singly-excited
manifold, so that the density operator reduces to

\[
\rho = \begin{pmatrix}
|0...0...0| & |10...0| & |0...1...0| & |0...0...1_k|
\end{pmatrix}
= \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & p_1 & 0 & 0 \\
0 & 0 & p_c & 0 \\
0 & 0 & 0 & p_k
\end{pmatrix}.
\]

(7)

For convenience, the case when none of the categories is chosen, element |0...0...0|, was included in the definition in Eq. (7) so that the density operator has dimensions \((k + 1) \times (k + 1)\). To perform the reduction of categories, the partial trace technique [20] is applied over the \(j^{th}\)-category to be reduced. This procedure yields a new density operator, \(\hat{\rho}_{j} = \text{tr}_j \hat{\rho}\), of dimension \(k \times k\) that is to be replaced, e.g., in equations (2-5) to calculate the change in the variability. The normalization factor needs to be replaced appropriately, specifically, it is replaced in favor of \(\hat{\rho}_{\text{max}} = \text{tr}_j \hat{\rho}_{\text{max}}\).

Due to the intrinsic diagonal character of \(\hat{\rho}\), the reduced density matrix \(\hat{\rho}_j\) will have all the information associated with the elements that it keeps and also information about the effects of the traced out elements (see, e.g., Sec. 2.2 in Ref. [20]). This characteristic is fundamental in applying the reduction of categories in a way consistent with information science [20].

The application of the partial trace for the present situation is straightforward. Since \(\rho\) is diagonal, performing the partial trace on the state |0...1...0| yields a reduced density matrix of order 2\(^k\) with the same elements of the total density matrix \(\rho\). The proportion \(p_j\) associated with the \(j^{th}\)-category switches to the none answer position, i.e., in the first input of the matrix. If the process is repeated over different states, the result is adding up the proportion associated with the traced element to the none answer position. Therefore, applying the partial trace over \(n\) of the \(k\) categories produces a reduced density operator \(\hat{\rho}_{\{j,...,n\}}\) of order \(2^{k-n+1}\) given by
After applying the reduced density operator to the equations (2-5), where \( n \) category reduction is performed, the following relatively simple expressions, in terms of the probabilities \( \{p\} \), account for measures of variability

\[
\sigma_L = 1 - \frac{1}{2(k-n-1)} \sum_{b=1}^{k} \left( \sum_{a \in R} p_a - p_b \right) + \sum_{b' = 1}^{k} \left| p_b - p_b' \right| \quad (9)
\]

\[
\sigma_S = 1 - \frac{1}{2(k-n-1)} \left( \sum_{b=1}^{k} \left( \sum_{a \in R} p_a - p_b \right)^2 \right)^{\frac{1}{2}} + \left( \sum_{b, b' = 1}^{k} \left( p_b - p_b' \right)^2 \right)^{\frac{1}{2}} \quad (10)
\]

\[
\sigma_P = \frac{\sum_{a \in R} p_a \log_2 (\sum_{a \in R} p_a) + \sum_{b=1}^{k} p_b \log_2 p_b}{\log_2 (k-n)} \quad (11)
\]

where \( R \) is the set of index associated with the reduced categories i.e. \( R = \{j, l, ..., n\} \). Not that that the number of categories have been reduced from \( k \) to \( k - n \).

Figure 5 depicts the results obtained for \( \sigma_P \) (top) and \( \sigma_S \) (bottom), applying the reduction of categories to the same dataset of the Problem-solving courts. From the initial ten categories scenario, the results of reduction of up to three categories are presented. Those results were obtained using the partial trace method (continuous line) and the Wilcox method (dashed line).

The reduction of categories is applied from the right to the left in Table 1 of Ref. [32]. As it can be seen from the figures, for all the territories in the U.S., the reduction of categories using the partial trace methods, implies an increase in the values of \( \sigma_S \) and \( \sigma_P \), conversely, a reduction of the values is obtained if the implemented method is the proposed by Wilcox. In terms of information, this is a very important issue, because the variability measures [Eqs. (4) and (5)] quantifies how the values between the categories are distributed, but in terms of information it can be understood as: if the value of variability approach to zero we have more information about organization between the categories and on the other hand, if the value approaches to 1 the lack of information information increases. Hence, when the reduction of categories is applied, information is lost, then it is natural to expect that the value of the variability increases instead of decreasing.

No that the variability for D.C. the increases when the categories are reduced using the Wilcox method; apparently, contradicting the conclusion above. This increase is due to the explicit value of the proportion of each reduced category. In this particular case, the first category reduced was other and for D.C. has a very high proportion of 0.727. Hence, ignoring that category, as in the Wilcox method, has a tremendous impact on the values of \( \sigma_S \) and \( \sigma_P \), otherwise, the method proposed by us keeps that information and thus do not produce a significant alteration in the curves.
Two main contributions have been done with this research: (i) Introducing the Hilbert space formalism and its advantages to solve reduction of categories problem of qualitative variables and (ii) Proposing a robustness analysis as a methodology to choose the best qualitative variation for a dataset. In more detail:

(i) Using the Hilbert space formalism, the matter of reduction of categories, an issue unsolved since the late 70’s is solved in a simple and very elegant way by performing the partial trace over the categories wanted to reduce. Importantly, this approach is consistent with information science. This allows for the extension of analysis without the necessity of a constructing a new dataset which sometimes is one of the biggest problems faced by social scientists. The formalism also allows the manipulation of datasets with simultaneous choice options.

(ii) A purely numerical methodology to choose the best variability measurement to implement in a dataset is new and provides a strong criteria if the dataset could have any bias or noise, like most of the real-datasets. The robustness against noise is a very important characteristic for indices, due to the fact that the intention of the index is to be useful and as general as possible.

The potential of the present formalism goes beyond the application of physics concepts and tools to social sciences and reach the field of Artificial Intelligence by providing formal support to the One Hot Encoding Algorithm [26].

VI. ACKNOWLEDGMENTS

Juan D. Botero was supported by Colciencias through a doctoral scholarship (Program No. 727)

[1] Simpson, E. Measurement of diversity. Nature 163, 688 (1949).
[2] Kvlseth, T. O. Note on biological diversity, evenness, and homogeneity measures. Oikos 62, 123–127 (1991). URL http://www.jstor.org/stable/35445460.
[3] Mokinski, F., Sheng, X. S., & Yang, J. Measuring disagreement in qualitative expectations. Journal of Forecasting 34, 405–426 (2015). For.2340.
[4] Wilcox, A. R. Indices of qualitative variation and political measurement. The Western Political Quarterly 26, 325–343 (1973). URL http://www.jstor.org/stable/446831.
[5] Les, M. & Maher, C. Measuring diversity: Choice in local housing markets. Geographical Analysis 30, 172–190 (1998).
[6] Mcdonald, D. G. & Dimnick, J. The conceptualization and measurement of diversity. Communication Research 30, 60–79 (2003).
[7] Bedeian, A. G. & Mossholder, K. W. On the use of the coefficient of variation as a measure of diversity. Organizational Research Methods 3, 285–297 (2000).
[8] Biemann, T. & Kearney, E. Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. Organizational Research Methods 13, 582–599 (2010).
[9] Kader, G. D. & Perry, M. Variability for categorical variables. Journal of Statistics Education 15, null (2007).
[10] Reynolds, H. Analysis of Nominal Data. No. nos. 4-7 in 07 (SAGE Publications, 1984). URL https://books.google.ca/books?id=loZlcoAek34C.
[11] Freedman, D., Pisani, R. & Purves, R. Statistics. International student edition (W.W. Norton & Company, 2007). URL https://books.google.com.au/books?id=mviJQgAACAAJ.
[12] Waldman, L. K. Types and measures of inequality. Social Science Quarterly 58, 229–241 (1977). URL http://www.
[13] Blau, P. M. *Inequality and heterogeneity: A primitive theory of social structure*, vol. 7 (Free Press New York, 1977).

[14] Agresti, A. & Agresti, B. F. Statistical analysis of qualitative variation. *Sociological Methodology* 9, 204–237 (1978). URL http://www.jstor.org/stable/270810.

[15] Teachman, J. D. Analysis of population diversity. *Sociological Methods & Research* 8, 341–362 (1980).

[16] McGrath, J. E., Berdahl, J. L. & Arrow, H. Traits, expectations, culture, and clout: The dynamics of diversity in work groups. (1995).

[17] Lau, D. C. & Murnighan, J. K. Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review* 23, 325–340 (1998).

[18] Williams, K. Y. & O’Reilly III, C. A. Demography and. *Research in organizational behavior* 20, 77–140 (1998).

[19] Harrison, D. A. & Klein, K. J. What’s the difference? diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review* 32, 1199–1228 (2007).

[20] Breuer, H. & Petruccione, F. *The Theory of Open Quantum Systems* (OUP Oxford, 2007). URL https://books.google.co.uk/books?id=DkcJPwAAQBAJ.

[21] Young, N. *An Introduction to Hilbert Space*. Cambridge mathematical textbooks (Cambridge University Press, 1988). URL https://books.google.com.co/books?id=yUycOJ5tQsUC.

[22] Berberian, S. *Introduction to Hilbert Space*. AMS Chelsea Publishing Series (AMS Chelsea, 1999). URL https://books.google.com.co/books?id=yUycOJ5tQsUC.

[23] Wootters, W. K. A wigner-function formulation of finite-state quantum mechanics. *Annals of Physics* 176, 1 – 21 (1987).

[24] Klimov, A., Romero, J. & Bj G. Discrete phase-space structure of n-qubit mutually unbiased bases. *Annals of Physics* 324, 53 – 72 (2009).

[25] Mokinski, F., Sheng, X. S. & Yang, J. Measuring disagreement in qualitative expectations. *Journal of Forecasting* 34 (2015).

[26] Müller, A. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O’Reilly Media, 2016). URL https://books.google.com.co/books?id=vbQ1DQAAQBAJ.

[27] Brualdi, R. *Combinatorial Matrix Classes*. No. v. 13 in Combinatorial matrix classes (Cambridge University Press, 2006). URL https://books.google.com.co/books?id=xdP9dSS1BxQC.

[28] Tsui, A. S., Egan, T. D. & O’Reilly III, C. A. Being different: Relational demography and organizational attachment. *Administrative science quarterly* 549–579 (1992).

[29] Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 3–55 (2001).

[30] Senders, V. L. Measurement and statistics: a basic text emphasizing behavioral science applications. (1958).

[31] Roehner, B. *Driving Forces in Physical, Biological and Socio-economic Phenomena: A Network Science Investigation of Social Bonds and Interactions* (Cambridge University Press, 2007). URL https://books.google.com.co/books?id=0PuzmiOlu10C.

[32] Strong, S. M., Rantala, R. R. & Kyckelhahn, T. Census of problem-solving courts, 2012. *US Department’s Bureau of Justice Statistics* NCJ 249803 (2016). URL https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=271950.