The complexity of financial wellness: examining survey patterns via kernel metric learning and clustering of mixed-type data

Jesse S. Ghashti
jesse.ghashti@ubc.ca
University of British Columbia
Kelowna, British Columbia, Canada

John R. J. Thompson
john.thompson@ubc.ca
University of British Columbia
Kelowna, British Columbia, Canada

ABSTRACT

Recent market events and inflation have significantly affected the financial stress facing many individuals, but understanding the main stressors is paramount to supporting them in making better long-term financial decisions. Financial advisors must understand the types of stress their clients face to provide tailored advice. While recent high inflation rates may underpin the cause of their clients' stress, we ask: what are the major sources of stress that affect an individual's financial wellness? In this study, we analyze the responses of 1874 individuals to 68 mixed-type questions from 2022 using distance-based clustering that is widely used in finance to group data into similar groups. Distance-based clustering is widely used in finance to group data into similar groups, which requires a predefined distance measurement between data points based on their (dis)similarity. We use a mixed-type metric that utilizes a variable-specific kernel functions with cross-validated bandwidths to optimally balance variables important for similarity, and smooth out variables irrelevant to the difference between data points. Applying the metric to the high-dimensional survey, we found two clusters of respondents: (1) the 'steady savers', who represent approximately one third of survey respondents and expressed stronger financial well-being with respect to day-to-day financial obligations and future outlooks, and (2) the 'financial strivers' who currently find themselves in more financially stressful situations. This segmentation provides financial advisors with useful results to allocate products, services, or advice tailored to support each group's unique financial wellness needs. By leveraging this methodology, we strive to advance the realm of personalized financial advising and the landscape of robo-advising. Enhanced precision and tailored strategies allow this work to elevate the quality of investment recommendations, contributing to the future of automated financial guidance.

CCS CONCEPTS

• **Information systems** → **Clustering**; • **General and reference** → **Empirical studies**; **Surveys and overviews**; **Metrics**; • **Computing methodologies** → **Cluster analysis**; • **Theory of computation** → **Unsupervised learning and clustering**; • **Mathematics of computing** → **Nonparametric statistics**.

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11.
https://doi.org/10.1145/3604237.3626849

KEYWORDS

Financial wellness, Survey data analysis, Mixed-type data, Metric learning, Kernel smoothing, Similarity, Distance-based clustering

ACM Reference Format:
Jesse S. Ghashti and John R. J. Thompson. 2023. The complexity of financial wellness: examining survey patterns via kernel metric learning and clustering of mixed-type data. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604237.3626849

1 INTRODUCTION

Clustering is a type of unsupervised learning that categorizes data points into distinct clusters based on their similarities, without the need for prespecified relationships or structures, allowing for the discovery of natural patterns and relationships that might not be apparent through traditional statistical models [21]. Clustering offers a blend of statistics and artificial intelligence for financial data analysis by offering a data-driven approach to identify patterns, group similar objects, and reveal hidden structures [33]. By organizing financial data (such as stocks, customers, and transactions) into meaningful clusters based on their similarities and dissimilarities, clustering methodologies enable financial institutions, investors, and analysts to target specific customer segments for accurate risk assessments, investment strategies, and tailored financial services.

The application of clustering in finance is diverse and multifaceted. Clustering enhances fraud detection by identifying anomalous patterns and suspicious activities [31]. Clustering also enables market segmentation by grouping customers with similar purchasing behaviors or risk preferences, facilitating targeted marketing campaigns and personalized financial products [30]. Clustering techniques can also be used in credit scoring [18], asset pricing [3, 13], credit risk assessment [22], and other areas where identifying patterns and segmenting data is needed for decision-making (see, e.g., [9, 36]).

Within clustering applications, diverse methodologies exist to address the challenge of handling mixed-type data. These approaches encompass discretization or numerical coding techniques to ensure uniformity of variable types, as well as the utilization of distance metrics specifically designed to accommodate mixed-type data. Discretization is a common strategy in machine learning that involves the conversion of continuous variables into categorical variables by employing domain knowledge-based interval specifications [10]. Following discretization, all variables are treated as categorical, enabling the selection of an appropriate clustering algorithm tailored explicitly for categorical variables, such as the k-Modes algorithm [19]. Such an approach may result in inaccurate interval specification, and may result in a substantial loss of information.

Conversely, numerical coding involves transforming categorical variables into numeric variables, enabling utilization of clustering methods applicable to continuous data, such as the k-means algorithm [16]. As suggested by [10], it is often challenging to assign reasonable values to categorical variables; instead, utilization of dummy coding techniques is required [25]. This approach generates a significantly higher-dimensional dataset, potentially leading to substantial ramifications for the clustering analysis. Unless the practitioner possesses extensive domain expertise, it is advised to refrain from employing either of these strategies in practice: instead, it is recommended to adopt a distance metric that accommodates each variable type effectively without any data transformations.

Gower's distance [12] is a widely used hybrid distance function that enables the computation of distances between two vectors of equal length, and remains prevalent in financial survey data clustering [26]. This metric incorporates a weighted combination of continuous and categorical distances. The categorical distance is determined based on the presence or absence of category matches, while the interval distance is scaled according to the variable range. However, the user-selected weights assigned to each variable may yield intractable solutions and produce varying results based on the dataset. Moreover, the logical interpretation of the simple matching coefficient for categorical variables is intuitively sound for binary variables but becomes less meaningful as the number of levels increases or when the categorical variable possesses an ordinal nature.

An alternative hybrid distance technique is the k-prototypes algorithm [19], which shares similarities with Gower's distance but utilizes a squared Euclidean distance for continuous variables and has also been used to cluster financial data [37]. Unlike Gower's distance, k-prototypes does not require variable-specific weights; instead, a single weight is applied to the entire categorical contribution of the distance function. To extend Gower's general coefficient of similarity to ordinal variables, the Podani distance metric [29] has been introduced, and the Wishart metric [42] is akin to the Podani metric but employs the sample standard deviation for continuous variables instead of the range. A comprehensive overview of hybrid distance metrics is provided in [10]. recent studies [11] suggest that this array of mixed-type distance metrics may not be optimal or even suitable in many scenarios.

The remainder of the paper is outlined as follows. In Section 2 we formalize the notion of the kernel distance for mixed-type data and define the proposed metric and algorithm. Section 3 describes the current study and data, along with the preprocessing steps. Section 4 describes the results of implementing the proposed metric in the context of two common clustering algorithms, namely, agglomerative hierarchical clustering and a modified k-means [7] algorithm, and provides a substantive clustering analysis. Section 5 concludes with avenues and suggestions for future research.

1.1 Our Contributions

This paper utilizes a kernel distance metric for mixed-type data, which offers a high degree of data-driven customization while minimizing the need for extensive preprocessing of financial data. By examining optimally selected bandwidths, practitioners can discern the relevancy of the variables in the dataset relative to one another for metric calculations. We emphasize the versatility of the resulting distances obtained from our proposed metric. This matrix can be integrated into any clustering technique that relies on distance matrices for clustering, which is useful in scenarios where only the distance matrix is available, and the original data is inaccessible. To validate the effectiveness of our proposed metric, we apply it to a high-dimensional financial wellness survey dataset. We employ well-established clustering methods, namely k-means and agglomerative hierarchical clustering, to uncover meaningful patterns within the data.

The application of the proposed metric to survey data, in conjunction with the clustering algorithms mentioned above, yields distinct profiles of two significant groups, unveiling noteworthy differences pertaining to their financial behaviours and attitudes. These clusters are denoted as "financial strivers" and "steady savers," and are defined by their distinctive characteristics, as discerned from the cluster prototypes. The resulting partitions not only highlight the heterogeneous financial profiles within these clusters, but also underscore the influence of their attitudes and behaviours on multifarious facets of workplace performance, stressors, as well as their perspectives on contemporary and future financial, political, and socioeconomic issues.

2 MIXED-TYPE KERNEL DISTANCE

2.1 Kernel Density Estimation and Bandwidth Selection

Kernel functions map data to a higher-dimensional space, while calculations happen in the original input space. This methodology is known as the "kernel trick" [34], which avoids the need to explicitly compute the coordinates of the data points in the higher-dimensional space. By applying kernel functions, data is transformed to make it easier to find patterns and make accurate predictions [5]. This paper uses kernel functions to calculate the similarities between observations within a mixed-type dataset. The kernel function can be used to define a similarity function between data points, which can be extended to define a distance metric for mixed-type data [11, 28].

Let $X^{n\times p}$ be a dataset of n observations and p variables. Denote $\mathcal{S} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ as an arbitrary similarity function with the following two properties: for any observation $\mathbf{x}_i$, $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_i) = 1$, and as the difference between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ increases, $\mathcal{S}(\mathbf{x}_i, \mathbf{x}_j)$ decreases. The similarity $\mathcal{S}$ can be cast as any symmetric kernel function satisfying these properties. We denote the kernel functions specific to datatypes as K, L, and ℓ for continuous (c), unordered (u) and ordered categorical variables (o), respectively. Denote bandwidths associated with each kernel function as $\boldsymbol{\lambda} \equiv \{\lambda^c, \lambda^u, \lambda^o\}$ where $\lambda^c \equiv \{\lambda_i\}_{i=1}^{p_c}$, $\lambda^u \equiv \{\lambda_i\}_{i=p_c+1}^{p_c+p_u}$, and $\lambda^o \equiv \{\lambda_i\}_{i=p_c+p_u+1}^{p}$. Note that $p = p_c + p_u + p_o$. We utilize the Gaussian kernel for (c), the Aitchison and Aitken kernel for (u) [2], and the Wang and van Ryzin kernel for (o) [40].

Before calculating a kernel distance metric, the optimal values of $\boldsymbol{\lambda}$ must be selected. A mixed-type joint kernel function between a vector $\mathbf{x}$ and an arbitrary point $\mathbf{x}_i = \{\mathbf{x}_i^c, \mathbf{x}_i^u, \mathbf{x}_i^o\}$, as defined in

[23], is written as

$$\mathcal{S}_{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{x}_i) = \prod_{k=1}^{p_c} \frac{1}{\lambda_k^c} K\left(\frac{x_k^c - x_{i,k}^c}{\lambda_k^c}\right) \times \ldots \\ \ldots \times \prod_{k=1}^{p_u} L\left(x_k^u, x_{i,k}^u, \lambda_k^u\right) \prod_{k=1}^{p_o} \ell\left(x_k^o, x_{i,k}^o, \lambda_k^o\right). \quad (1)$$

Optimal bandwidth selection methods are designed to preserve estimator convergence while having several other desirable properties, including smoothing out irrelevant variables [24]. There is a wide range of methods for optimal bandwidth selection (see [14, 20, 32, 35]). In this paper, we use maximum-likelihood cross-validation (MLCV) through the *R* package *np* for our bandwidth selection criterion [17], where the MLCV objective function to be minimized is

$$CV(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \ln\left(\frac{1}{(n-1)} \sum_{j=1, j\neq i}^{n} \mathcal{S}_{\boldsymbol{\lambda}}(\mathbf{x}_i, \mathbf{x}_j)\right) \\ = \sum_{i=1}^{n} \ln\left(\hat{\mathcal{S}}_{-i}(\mathbf{x}_i)\right), \quad (2)$$

where $\hat{\mathcal{S}}_{-i}(x_i)$ is the leave-one-out estimator of $\mathcal{S}_{\boldsymbol{\lambda}}(\cdot)$ in Equation (1). Optimal bandwidth selection that maximizes similarity is a crucial step in metric learning, as a naïve choice can have adverse effects on clustering performance. Bandwidths have a well-defined range of values, with $0 < \lambda^c < \infty$, $0 \leq \lambda^u \leq \frac{c-1}{c}$, and $0 \leq \lambda^u \leq 1$, where c is the number of categories of the p_uth categorical variable. If the estimated λ of the pth variable approaches their upper bound, the variable is effectively smoothed from the data and does not contribute to the overall distance. A low bandwidth value means the kernel has a narrow width, resulting in a higher concentration of influence around each data point, which we can use to determine variable importance.

2.2 Kernel Density Sum (KDSUM) Distance and Algorithm

The pairwise similarity between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$\psi(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\lambda}) = \prod_{k=1}^{p_c} \frac{1}{\lambda_k^c} K\left(\frac{x_{i,k}^c - x_{j,k}^c}{\lambda_k^c}\right) + \ldots \\ \ldots + \sum_{k=p_c+1}^{p_u} L(x_{i,k}^u, x_{j,k}^u, \lambda_k^u) + \sum_{k=p_c+p_u+1}^{p} \ell(x_{i,k}^o, x_{j,k}^o, \lambda_k^o). \quad (3)$$

By the definition of the kernel and similarity functions, $\psi(\cdot)$ satisfies the similarity properties [6] and is a similarity function, and we can set $\mathcal{S}(\cdot) := \psi(\cdot)$. Combining the similarity properties, and adapting the kernel distance described by [28] to the multivariate setting, we define the distance between any two data points $\mathbf{x}_i$, $\mathbf{x}_j$ of the dataset X as

$$d(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\lambda}) = \psi(\mathbf{x}_i, \mathbf{x}_i|\boldsymbol{\lambda}) + \psi(\mathbf{x}_j, \mathbf{x}_j|\boldsymbol{\lambda}) - 2\psi(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\lambda}) \\ = 2(1 - \psi(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\lambda})). \quad (4)$$

Using the properties of similarity [6], one can easily show that Equation 4 is a well-defined distance metric. Algorithm 1 provides the pseudocode for implementing the KDSUM metric.

Algorithm 1 KDSUM

1: Given a dataset X, assign variable types and order as $[p_c, p_u, p_o]$
2: Select symmetric kernel functions K, L, ℓ.
3: Calculate optimal bandwidths for each p_i using cross-validation procedure using Equations (1) and (2) for selected kernels in Step 2
4: Calculate the pairwise distance between all observations $\mathbf{x}_i$ and $\mathbf{x}_j$ using Equations (3) and (4) and the selected kernels in Step 2 to obtain the dissimilarity matrix with bandwidths from Step 3
5: Cluster dissimilarity matrix with any clustering algorithm that accepts the matrix as input

3 STUDY DESCRIPTION

The anonymized 2022 National Payroll Institute's annual financial survey financial wellness survey was completed voluntarily amongst employed individuals, covering various aspects of their profiles, employment, remuneration, financial situation, financial literacy, saving, debt, economic confidence, pay and tax statements, payroll/tax/deductions/benefits, etc. The survey length varies annually, and does not include individuals in school, retired, or unemployed, which may result in differing experiences among specific groups. For the purposes of this paper, we cluster the 2022 survey, where the original dataset consists of 68 questions and 1874 participants.

For questions that allowed multiple selections (e.g., "select all that apply"), a binary representation approach is adopted, where each option within these questions was partitioned into a separate column, with a value of 1 indicating participant selection. Several questions involving Likert-scale type ratings contained multiple sub-questions presented together: these sub-questions were partitioned into individual variables. Missing values in the dataset represented cases where participants either did not select a specific option or did not answer the question at all. These missing values were imputed as zeros. Text-entry only questions, which required Natural Language Processing (NLP) for grouping similar answers, were excluded from the analysis, and are an area of future work (see [38] for applications of NLP in mutual fund categorization). In cases where participants selected the "Other (please specify)" option for certain questions, the corresponding text-entry columns were removed to avoid reliance on NLP. It was determined that removing these columns would not significantly affect the overall information captured, since the responses were already recorded in the previous numeric column. To ensure data quality, participants with over 100 "zero" responses were excluded from the analysis. This step aimed to filter out incomplete or low-effort surveys, removing 32 participants and leaving 1842 remaining participants. Following data cleaning, the dataset consisted of 186 remaining columns.

The data were categorized based on the question types. For example, "Select all that apply" questions were treated as unordered

Table 1: Five lowest bandwidths selected through maximum-likelihood cross-validation for unordered and ordered categorical variables, along with the question type.

Unordered		Ordered	
Question	Type	Question	Type
Q46	Stress	Q27	Spending
Q1	Socio-demographic	Q51	Debt
Q48	Debt	Q13	Stress
Q47	Debt	Q11	Costs
Q40	Retirement	Q52	Financing

categorical variables, while Likert-type scale questions were treated as ordered categorical variables. Based on intuition, the remaining questions were carefully categorized into their respective variable types, resulting in a total of 2 continuous, 110 unordered categorical, and 74 ordered categorical variables.

Sample questions of each variable type:

Continuous: "Q18: How long is your average commute time on a typical workday (one way)?" The response options ranged from 1 minute to 300 minutes.

Unordered categorical: "Q62: Which of the following best describes your family status?" The response options included: 1) Married/partnered with children, 2) Married/partnered without children, 3) Single/separated/divorced/widowed with children, 4) Single/separated/divorced/widowed without children, and 5) I prefer not to answer.

Ordered categorical: "Q28: How frequently or infrequently do you budget your expenditures?" The response options ranged from 1) Very frequently to 7) Very infrequently, and 8) I do not know / Not applicable.

4 RESULTS

4.1 Variable Relevancy via Optimal Bandwidth Selection

Using the algorithm described in Section 2.1, the two continuous variables attained bandwidths $(\lambda_1^c, \lambda_2^c) = (26.12, 39.52)$, for questions 18 and 19, respectively using a Gaussian kernel. The bandwidths for λ^u and λ^o are visualized in Figure 1 using the Aitchison & Aitken and Wang & van Ryzin Kernel, respectively. From the figure, one can observe that the majority of the λ_i^o are greater than $0.5 \times \max(\lambda_i^o)$, indicating many of these variable types are being smoothed from the data, whereas many of the $\lambda_i^u \ll 0.5 \times \max(\lambda_i^u)$, implying a larger contribution to the overall distance calculation. As a ratio of the observed and maximum bandwidths, five of each question type (ordered and unordered categorical) associated with the lowest bandwidths ratios are shown in Table 1, indicating they contribute most heavily to the distance calculation. The table shows that questions related to employment status, current debt, and paying off debt are considered most important. The specific questions can be found in Appendix A.

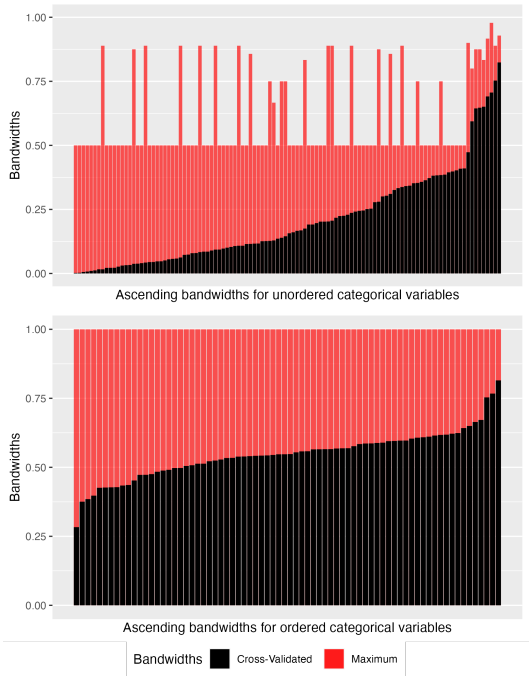

Figure 1: Maximum-likelihood cross-validated bandwidths in comparison to their upper bounds for the unordered and ordered categorical variables

4.2 Optimal Number of Clusters

The flexibility of KDSUM makes it adaptable to any clustering algorithm that accepts a distance matrix as input, rather than the dataset itself. Recent studies suggest that Agglomerative Hierarchical Clustering Techniques (see, e.g., [8, 27]) perform best with this type of metric. We implement agglomerative hierarchical clustering and a modified $k-$Means algorithm [7], which accepts a distance matrix, and compare results.

Before clustering implementation, it is important to determine the optimal number of clusters using the distance matrix. We conduct a search using a magnitude of metrics in R package *NbClust* [4] for both clustering techniques described above. Both $k-$Means and hierarchical clustering (Ward's method [41]) were in agreeance amongst all indices, and the results are shown in Figure 2. Among the 22 indices considered, the most frequently occurring number of clusters was two, with three as the second most frequent. An exploratory analysis of the differences between two and $C = 3$ clusters did not provide any overwhelmingly intuitive differences in structure, so we chose two clusters for analysis. We note that the conclusions between the two clustering methods are very similar with respect to two and three clusters. In fact, for two clusters, the prototype for cluster 2 was the same for both clustering methods, and for cluster 1, the individuals between both methods were very similar. Therefore, we will agglomerate the clustering results and not distinguish between the two clustering approaches any further: rather, we will characterize the overarching relationships between clusters for two clusters.

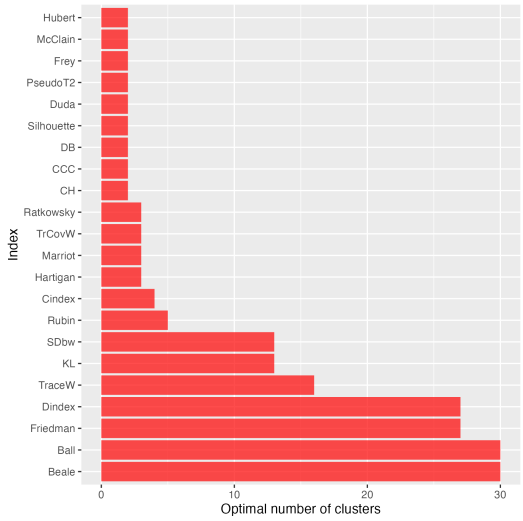

Figure 2: Optimal number of clusters amongst 22 indices. The most common was $C = 2$, and $C = 3$ was second in agreeance.

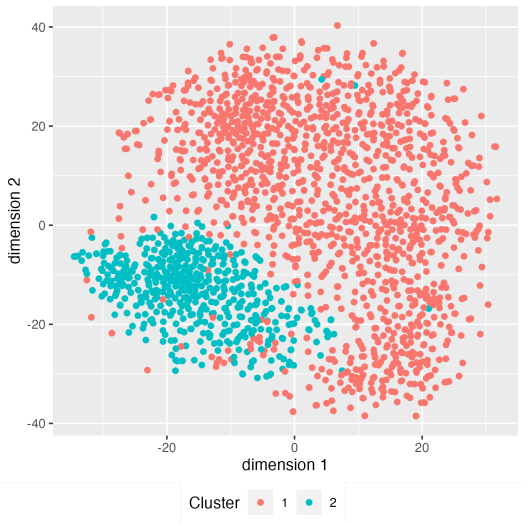

Figure 3: t-SNE plot for two clusters using the $k-$means clustering for distance matrices. The plot indicates two well-separated clusters, where C_1 (red) and C_2 (blue) represent 64% and 36% of respondents, respectively.

Post-processing, the cluster centers were analyzed via a cluster prototype. Since both algorithms have their own notion of a cluster center, manual implementation of a cluster prototype was determined by subsetting the individuals to each of their respective clusters for each algorithm, calculating the cluster mean (centroid), and then assigning the individual with the closest proximity to the cluster mean as a prototype for the cluster. This allows the practitioner to analyze the individual who is a central representative of the entire cluster and to determine any distinguishing differences between the clusters. Due to the high dimensionality (186 dimensions) of the dataset, it is impractical to list the complete prototype of each cluster; instead, we characterize the distinguishing features between individuals allocated to each cluster.

4.3 Clustering Results

Based on the distinctive characteristics exhibited by the medoids, we classify cluster 1 as the "financial strivers" and cluster 2 as the "steady savers." The financial strivers constitute approximately 64% of the individuals surveyed, while the remaining 36% are deemed the steady savers. The t-distributed stochastic neighbor embedding (t-SNE) [39], depicted in Figure 3, visualizes the two relatively well-separated clusters based on the distances calculated by mapping observations into a two-dimensional space.

The financial strivers comprise individuals who demonstrate an optimistic outlook on the future of finance and the economy. This group predominantly consists of non-homeowners who face longer commuting distances to their workplaces. A significant proportion (31-40%) of their household income is allocated towards housing costs, rendering their monthly expenses unaffordable. They have experienced job changes, such as resignations or job transitions, within the past year, and typically these individuals tend to spend their entire net pay without actively budgeting their monthly expenditures. However, they express a general financial goal of saving, and to contribute more towards their savings, they have curtailed expenses such as recreational activities, personal items, and hobbies. On average, they save approximately 16-20% of their income and actively strive to increase their savings compared to the previous year, with the primary motive for saving being to fund significant purchases, such as a house or a car. They estimate that retiring comfortably would require savings in the range of $600,001 to $700,000 and anticipate their retirement income to come from sources such as inheritance, other investments, or a potential lottery win.

The financial strivers frequently occupy positions in accounting, finance, or human resources within companies employing 100–199 individuals. They express concerns regarding the future state of the economy and geopolitical factors, particularly the decline in housing values, potential job loss, mounting debt, and the impact of inflation on retirement plans. The financial stress experienced by these individuals within their workplaces has resulted in diminished productivity, subpar performance evaluations, workplace accidents, and, in some cases, voluntary resignations. They commonly carry debt from home equity lines of credit, student loans, credit cards, and loans from family members. The necessity of making minimum debt payments often impedes their ability to save, and a delayed paycheque would pose some difficulty in meeting their financial obligations, where they would likely seek assistance from family or friends. Factors that could prompt these individuals to consider alternative employment opportunities include improved health benefits, retirement benefits, and educational funding.

Conversely, the steady savers constitute individuals employed in companies with 500–999 employees, typically in roles unrelated to accounting, finance, or human resources. This group demonstrates a more positive outlook regarding their financial situation and expresses confidence in their ability to fulfill their financial obligations, even if their paycheque were to be delayed. In such an event, they would most likely utilize their savings to ensure

timely bill payments. The primary motivation for considering a job change among the steady savers is higher wages. Although they experienced an average pay increase of 2% in the past year, they hold the belief that their income may not keep pace with inflation, and anticipate a rise in personal costs across various domains due to inflation. These individuals harbour concerns regarding the strength of the economy, future prospects, inflationary pressures, stock market fluctuations, the lingering effects of COVID-19, and geopolitical issues.

Similar to the financial strivers, the steady savers share concerns about declining house values, escalating housing expenses, interest rates, and job security. However, they exhibit a comparatively lower apprehension about their debt burden or retirement plans. The majority of the steady savers are homeowners who made their purchases more than two years ago, and have the financial means to allocate only 11-20% of their monthly household income towards housing costs, which they perceive as manageable.

Unlike financial strivers, steady savers tend to spend less than their net pay and actively budget their monthly expenditures. Their primary financial goal revolves around saving for retirement, and they believe increasing their income is the most effective strategy to achieve this objective. To contribute more to their savings, they have reduced expenditures related to personal travel, trips, and vacations. On average, they save approximately 11-15% of their income, and estimate that a comfortable retirement would require savings in the range of $500,001 to $600,000. The main sources of retirement income is expected to come from government-sponsored pension funds, employer-supported pension funds, and savings accumulated through Registered Retirement Savings Plans (RRSPs) and Tax-Free Savings Accounts (TFSAs). Their target retirement age is 60, with the possibility of postponing retirement if their investments do not yield the expected returns.

In contrast to the financial strivers, steady savers do not perceive their personal finances as having a negative impact on their workplace performance or their relationships with family and friends. They do not feel as overwhelmed by their debt burden and spend approximately 30–45 minutes of their workday managing financial matters, and financial stress has generally not caused any issues at work for this group. It is worth noting that the median response indicates that the majority of individuals among the steady savers do not carry any debt. Demographically, the steady savers are predominantly male, aged 50-59, married with children, residing in Eastern Canada, and have an annual household income ranging from $80,000 to $99,999.

The similarities between the clusters are:

- Both clusters express concerns about the strength of the economy, inflation, and the decline in the value of their houses.
- They have a financial goal of saving for retirement, although the specific strategies and savings amounts differ.
- Both clusters allocate a significant portion of their income towards housing costs.
- They are both working professionals with stable jobs, albeit in different company sizes and industries.
- Both clusters have concerns about future economic conditions, including geopolitical issues.
- The demographics between each cluster are similar. The key distinction between $C = 2$ and $C = 3$ clusters is primarily demographic, whereas the new cluster narrowly focused on the densely populated Greater Toronto Area, where commute times were significantly higher, and the stress levels related to debt and saving were closely aligned to the financial strivers.

Distinguishing differences between the clusters:

- Financial strivers are composed of non-homeowners with longer commutes, while the steady savers consist of homeowners with shorter commutes.
- Financial strivers struggle with affordability, while steady savers find their housing costs more manageable and affordable.
- Financial strivers tend to spend all of their net pay and do not usually budget their expenditures, while steady savers spend less than their net pay and actively follow budgets.
- Financial strivers have experienced job changes in the past year, while steady savers have not.
- Financial strivers express concerns about various financial stressors impacting workplace performance and relationships, while steady savers do not report such negative effects.
- Financial strivers have a higher level of debt, including home equity lines of credit, student loans, credit cards, and family loans, whereas steady savers have minimal or no debt.
- Financial strivers anticipate retirement income from inheritance, investments, or lottery wins, while steady savers expect income from pension funds and personal savings.

Overall, the two clusters demonstrate distinct financial characteristics, including their housing situations, spending habits, savings strategies, concerns, and demographics. From the clustering results, the distribution of individuals between the two clusters is most evident from four questions, as seen in Figure 4. The specific questions can be found in Appendix A.

4.4 Comparison to Other Mixed Distance Metrics

In comparing mixed-type distance metrics using a consistent modified $k-$means clustering algorithm for distance matrices, and without true class labels, we employ average Euclidean distance between standardized dataset cluster centers to ensure equal variable contributions. This enables assessing similarity among clustering outcomes based on the distance metric. Although using measures like within-group sum of squares is impractical due to varying scales of distance matrices and the clustering methodology's direct use of a distance matrix, we scale the matrices, recluster, and examine within-group sum-of-squares to evaluate algorithmic performance regarding cluster compactness and homogeneity, noting the scaling of distance matrix had no effect on the clustering results.

We compare the KDSUM metric to six mixed-type distance metrics, namely Gower's Distance (GOW) [12], Wishart (WIS) [42], Podani (POD) [29], Huang (HUA) [19], Harikumar-PV (HP) [15],

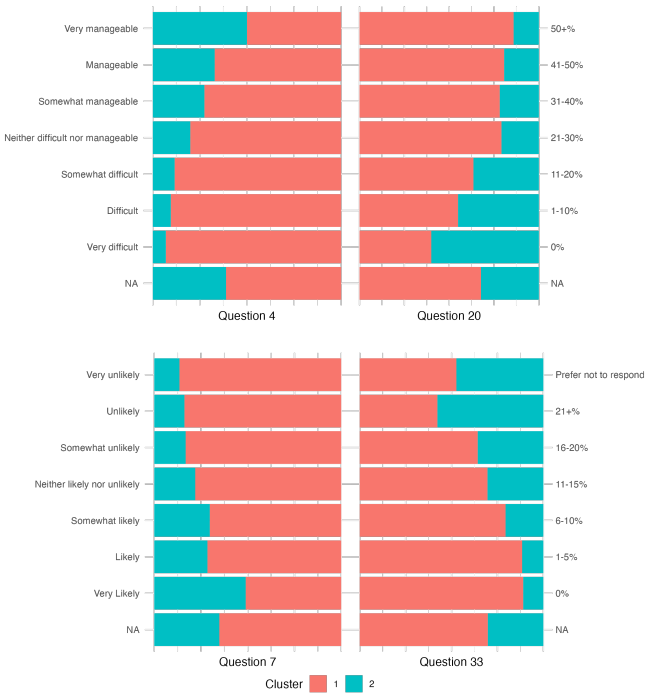

Figure 4: Four selected questions and the distribution of responses amongst the two clusters. The results align closely with the two defined clusters profiles.

and Ahmad-Dey (AD) [1], defined in Appendix B, with results in Table 2. The cluster centers of KDSUM are similar to those of Gower's, Wishart, and Podani, logically correlating with their distance metric structures. KDSUM aligns most with Wishart, while Huang, Harikumar-PV, and Ahmad-Dey exhibit mutual similarity, especially between Huang and Ahmad-Dey. The proposed methodology attains the most compact, homogeneous clusters, offering a 27.96% average improvement in the WSS total, with improvements ranging from 14.59% (Wishart) to 61.75% (Harikumar-PV). Note that the WSS for KDSUM for CL1, CL2, and Total were 263.6, 38.7, and 302.3, respectively.

5 CONCLUSION

This paper extends mixed-type financial data to higher dimensions for the purpose of a distance calculation, and then further clusters that data based on their similarities using agglomerative and modified $k-$Means clustering algorithms. Our study examined the financial behaviors and attitudes of employed individuals using survey data. Through a rigorous analysis process, we identified two distinct clusters among the participants: "financial strivers" and "steady savers."

The "financial strivers" exhibited several common concerns. They expressed worries about housing costs, job changes, and debt. This group likely faces financial challenges and may require assistance or guidance in managing their finances effectively. Understanding

Table 2: Comparison of the average Euclidean distance between cluster centers for various distance metrics, based on the modified $k-$means algorithm for a standardized version of the examined data. We also examined the within-group sum-of-squares (WSS) for each cluster (C1,C2) and the total WSS for each clustering method based on the standardized distance matrix.

	GOW	WIS	POD	HUA	HP	AD
KDSUM	1.174	0.719	0.934	2.813	2.871	2.815
GOW		0.989	0.995	2.439	2.448	2.403
WIS			0.728	2.626	2.651	2.628
POD				2.690	2.686	2.635
HUA					0.324	0.145
HP						0.293
WSS (CL1)	301.4	219.9	237.1	271.0	482.9	277.3
WSS (CL2)	103.1	134.0	138.5	101.2	306.4	97.0
WSS (TOT)	404.5	353.9	375.6	372.3	789.2	374.3

the specific concerns and stressors faced by this cluster can inform the development of targeted interventions and support programs to address their needs. The "steady savers" demonstrated a higher level of confidence in their financial situation, displayed active budgeting behaviors, and exhibited a more proactive approach to financial management indicating a greater likelihood of financial stability and preparedness for unexpected financial burdens. By identifying these distinct clusters, we contribute to a better understanding of the diverse financial situations and challenges faced by individuals in the workforce. Policymakers, financial institutions, and educators can utilize these findings to develop targeted strategies, educational programs, and resources to improve financial well-being and promote positive financial behaviors among employed individuals.

There are several promising directions for future research. The first step is to investigate optimal bandwidth selection procedures for kernel metrics utilized in various distance-based clustering algorithms. While agglomerative hierarchical clustering was selected for this study for ease of demonstration, a new or existing algorithm may further enhance the classification and clustering of mixed data with a kernel distance metric. A detailed analysis of clustering algorithms that require dissimilarity matrices as input and determining the optimal clustering algorithm that pairs with kernel distance metrics is also future work.

Considerations for metric-specific research directions include cross-validatory bandwidth selection and kernel function specification. While bandwidth choices can be arbitrary, substituting $\mathcal{S}_{\boldsymbol{\lambda}}(\cdot)$ in Equation (1) with the KDSUM metric from Equation (3) may maximize separation between observations and clusters. Preliminary research indicates the choice of kernel functions for the KDSUM metric minimally impacts performance when variable-specific kernel functions are used, though further analysis is needed. By far, the largest contribution of error for kernel choices comes from the misspecification of data type (e.g. unordered kernel in place of ordered).

A thorough analysis of the results from repeated financial wellness surveys conducted over multiple years is necessary. Participants frequently complete this survey on an annual basis, and utilizing kernel metric learning to evaluate and investigate the financial wellness of individuals over time in similar or changing market conditions can serve as a crucial step towards a more comprehensive examination of their financial well-being and behaviors. By monitoring the clustering patterns over time, we may gain improved insight into the behavioral trends and future actions of respondents belonging to specific clusters, as well as identify the potential emergence of new clusters. For instance, distinct trajectories could be observed among the "financial strivers" cluster over time, and this larger cluster of individuals could be further segmented into smaller clusters based on their evolving financial wellness. Discovering these smaller, well-defined clusters may be beneficial for financial advisors, as it would enable a more focused selection of products, services, and tailored advice to cater to the specific needs of individuals. Therefore, a serially dependent clustering methodology is considered for future work for analyzing clusters that can evolve over time. Additionally, we can delve deeper into the applicability and efficacy of the KDSUM method in clustering mixed-type data for Robo-advising and automated financial guidance.

ACKNOWLEDGMENTS

The authors would like to thank Canada's Financial Wellness Lab and the National Payroll Institute for providing the financial survey data analyzed in this research and acknowledge the funding support of the University of British Columbia and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] Amir Ahmad and Lipika Dey. 2007. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering* 63, 2 (2007), 503–527.
[2] John Aitchison and Colin G. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63, 3 (1976), 413–420.
[3] Giulia Caruso, SA Gattone, Francesca Fortuna, and Tonio Di Battista. 2021. Cluster Analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences* 73 (2021), 100850.
[4] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. NbClust: An R Package for Determining the Relevant number of clusters in a data set. *Journal of Statistical Software* 61, 6 (2014), 1–36. https://www.jstatsoft.org/v61/i06/
[5] Daoqiang Zhang Songcan Chen. 2002. Fuzzy clustering using kernel method. *IEEE, Nanjing, China* (2002).
[6] Shihyen Chen, Bin Ma, and Kaizhong Zhang. 2009. On the similarity metric and the distance metric. *Theoretical Computer Science* 410, 24-25 (2009), 2365–2376.
[7] Michael Christoph and Quirin Stier. 2021. Fundamental clustering algorithms suite. *SoftwareX* 13 (2021), 100642.
[8] William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification* 1 (1984), 7–24.
[9] Thomas J De Luca and Dhagash Mehta. 2023. ESG Fund Usage among Individual Investor Households: A Machine Learning–Based Behavioral Study. *The Journal of Impact and ESG Investing* 3, 3 (2023), 28–44.
[10] Alexander H. Foss, Marianthi Markatou, and Bonnie Ray. 2019. Distance metrics and clustering methods for mixed-type data. *International Statistical Review* 87, 1 (2019), 80–109.
[11] Jesse S. Ghashti and John. R. J. Thompson. 2023. Kernel Metric Learning for Clustering Mixed-type Data. *arXiv preprint arXiv:2306.01890* (2023).
[12] John C Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics* (1971), 857–871.
[13] Olesya V Grishchenko and Marco Rossi. 2012. The role of heterogeneity in asset pricing: The effect of a clustering approach. *Journal of Business & Economic Statistics* 30, 2 (2012), 297–311.
[14] Peter Hall. 1981. On nonparametric multivariate binary discrimination. *Biometrika* 68, 1 (1981), 287–294.
[15] Sandhya Harikumar and PV Surya. 2015. K-medoid clustering for heterogeneous datasets. *Procedia Computer Science* 70 (2015), 226–237.
[16] John A. Hartigan and Manchek A. Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.
[17] Tristan Hayfield and Jeffrey S. Racine. 2008. Nonparametric Econometrics: The np Package. *Journal of Statistical Software* 27, 5 (2008). https://www.jstatsoft.org/v27/i05/
[18] Nan-Chen Hsieh. 2005. Hybrid mining approach in the design of credit scoring models. *Expert systems with applications* 28, 4 (2005), 655–665.
[19] Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2, 3 (1998), 283–304.
[20] Clifford M. Hurvich, Jeffrey S. Simonoff, and Chih-Ling Tsai. 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60, 2 (1998), 271–293.
[21] Anil K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666.
[22] Gang Kou, Yi Peng, and Guoxun Wang. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information sciences* 275 (2014), 1–12.
[23] Qi Li and Jeffrey S. Racine. 2007. *Nonparametric econometrics: Theory and practice.* Princeton University Press.
[24] Clive R. Loader. 1999. Bandwidth selection: classical or plug-in? *The Annals of Statistics* 27, 2 (1999), 415–438.
[25] Brendan McCane and Michael Albert. 2008. Distance functions for categorical and mixed variables. *Pattern Recognition Letters* 29, 7 (2008), 986–993.
[26] Adam Metzler, Yuhao Zhou, and Chuck Grace. 2019. Learning about financial health in Canada. *Available at SSRN 3507769* (2019).
[27] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 1 (2012), 86–97.
[28] Jeff M. Phillips and Suresh Venkatasubramanian. 2011. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625* (2011).
[29] János Podani. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* 48, 2 (1999), 331–340.
[30] Girish Punj and David W. Stewart. 1983. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research* 20, 2 (1983), 134–148.
[31] Andrei S. Sabau. 2012. Survey of Clustering based Financial Fraud Detection Research. *Informatica Economica* 16, 1 (2012), 110–122.
[32] Stephan R. Sain, Keith A. Baggerly, and David W. Scott. 1994. Cross-validation of multivariate densities. *J. Amer. Statist. Assoc.* 89, 427 (1994), 807–817.
[33] Iqbal H. Sarker. 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science* 2, 5 (2021), 377–399.
[34] Bernhard Schölkopf. 2000. The Kernel Trick for Distances. In *Advances in Neural Information Processing Systems*, Vol. 13. MIT Press.
[35] Bernard W Silverman. 1986. *Density estimation for statistics and data analysis.* Vol. 26. CRC press.
[36] Fu Tan and Dhagash Mehta. 2022. Health state risk categorization: A machine learning clustering approach using health and retirement study data. *The Journal of Financial Data Science* 4, 2 (2022), 139–167.
[37] John RJ Thompson, Longlong Feng, R Mark Reesor, and Chuck Grace. 2021. Know Your Clients' behaviours: a cluster analysis of financial transactions. *Journal of Risk and Financial Management* 14, 2 (2021), 50.
[38] Dimitrios Vamvourellis, Mate Toth, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. 2022. Learning Mutual Fund Categorization using Natural Language Processing. In *Proceedings of the Third ACM International Conference on AI in Finance*. 87–95.
[39] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[40] Min-Chiang Wang and John van Ryzin. 1981. A class of smooth estimators for discrete distributions. *Biometrika* 68, 1 (1981), 301–309.
[41] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.
[42] David Wishart. 2003. K-means clustering with outlier detection, mixed variables and missing values. In *Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation eV (Exploratory Data Analysis in Empirical Research)*, Reginald N. Smythe and Alexander Noble (Eds.). Springer Berlin Heidelberg, University of Munich, 216–226.

A REFERENCED QUESTIONS

Below is a selection of questions that are referenced throughout this manuscript.

Q1: Please tell us your employment status: (Check all that apply.) *options: full-time, part-time seasonal (≤30/week), seasonal, self-employed, gig worker, retired, student, other, not employed*

Q4: If your paycheque (i.e., payment of salary or wages) was delayed for a week, how difficult would it be to meet your current financial obligations? *options: very difficult, difficult, somewhat difficult, neither difficult nor manageable, somewhat manageable, manageable, very manageable, I do not know / not applicable (NA)*

Q7: How likely are you to come up with $20,000 if an emergency arose within the next month? *options: very likely, likely, somewhat likely, neither likely nor unlikely, somewhat unlikely, unlikely, very unlikely, I do not know / not applicable (NA)*

Q11:For each of the following types of expenses, indicate whether you think your spending will increase, decrease or stay the same over the next year (due to inflation). *options: housing costs, transportation costs, food costs, utility costs, medical or healthcare costs, childcare costs, personal costs, clothing costs, entertainment/ membership/ subscription costs, travel costs, discretionary costs, other*

Q13: When thinking about things beyond an individual's control, how concerned are you about each of the following issues? *options: the strength of the economy (today), outlook for the economy, inflation/increases in cost of living, the Stock Market, the COVID-19 pandemic, geopolitical issues*

Q20: Approximately what portion of your monthly Household Income is typically consumed by your total monthly housing costs? (Including mortgage or rent, utilities, property taxes, insurance, and maintenance) *options: none (0%), 1-10%, 11-20%, 21-30%, 31-40%, 41-50%, 50+%, I do not know / not applicable (NA)*

Q27: Which of the following tends to be true for you in a typical or average pay period? *options: I spent more than my net pay, I spend all my net pay, I spend less than my net pay, I do not know / not applicable (NA)*

Q33: On average, what percentage of your paycheque do you put toward savings? *options: 0%, 1-5%, 6-10%, 11-15%, 16-20%, 21+%, I prefer not to respond, I do not know / not applicable (NA)*

Q40: How do you plan to finance your retirement? Indicate all the potential sources of funds that you intend to use to fund your retirement. (Choose all that apply.) *options: government-sponsored pension funds, government-sponsored senior citizen income support programs, employer supported pension funds, inheritances or intergenerational wealth, savings (through RRSPs), savings (through TFSAs), savings (through other investments), selling an asset, lotter win, other, I do not know / not applicable (NA)*

Q46: Has personal financial stress caused any of the following to happen to your life at work? (Check all that apply.) *options: caused me to take a personal or sick day or leave, decreased my motivation at work, resulted in a decrease in productivity at work, resulted in a poor performance review, resulted in a workplace accident or safety concern, resulted in me leaving my job or seeking other employment, resulted in me requesting a part-time arrangement or a reduction in hours, resulted in strained relationships or interactions with colleagues, other, none of the above, I do not know / not applicable (NA)*

Q47: What type(s) of debt do you currently have? (Clease check all that apply.) *options: mortgage(s) on my principal residence, mortgage(s) on a rental or business property, line of credit, home equity line of credit, student loan, car loan, payday loans, consolidation loans, credit card debt, family debt, other debt (specify), I do not have debt, I prefer not to respond.*

B MIXED-TYPE DISTANCES

Below are the formulas for mixed-type distance measures considered in the application for this paper. Let R_k be the range of the variable k and s_k be the standard deviation of a variable k, p be the total number of variables (p_n for continuous, p_c for categorical), and $\omega_{ijk} \in \{0, 1\}$ for missing values.

Gower: $d_{ij} = 1 - s_{ij}$, where

$$s_{ij} = \frac{\sum_{k=1}^{p} \omega_{ijk} s_{ijk}}{\sum_{k=1}^{p} \omega_{ijk}}, \quad s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}.$$

Wishart:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \omega_{ijk} \left(\frac{x_{ik} - x_{jk}}{\delta_{ijk}} \right)^2},$$

where $\delta_{ijk} = s_k$ if k is continuous, and $\delta_{ijk} \in \{0, 1\}$ if categorical.

Podani:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \omega_{ijk} \left(\frac{x_{ik} - x_{jk}}{\delta_{ijk}} \right)^2},$$

where $\delta_{ijk} = R_k$ if k is continuous, and $\delta_{ijk} \in \{0, 1\}$ if categorical.

Huang:

$$d_{ij} = \sum_{k=1}^{p_n} (x_{ik} - x_{jk})^2 + \frac{\sum_{k=1}^{p_n} s_k^2}{p_n} \sum_{l=1}^{p_c} \delta_c(x_{il} - x_{jl}),$$

where $\delta_c(\cdot)$ is the simple matching distance.

Harikumar-PV:

$$d_{ij} = \sum_{k=1}^{p_n} |x_{ik} - x_{jk}| + \sum_{l=1}^{p_c} \Delta_c(x_{il} - x_{jl}) + \sum_{m=1}^{p_b} \Delta_b(x_{im}, x_{jm}),$$

where p_b is the number of binary variables, $\Delta_c(\cdot)$ is the co-occurrence distance, and $\Delta_b(\cdot)$ is the Hamming distance.

Ahmad-Dey:

$$d_{ij} = \sum_{k=1}^{p_n} (x_{ik} - x_{jk})^2 + \sum_{l=1}^{p_c} \Delta_c(x_{il} - j_{il}),$$

where $\Delta_c(\cdot)$ is the co-occurrence distance.