DEPOCEN
Working Paper Series No. 2011/18

# Informality, business outcomes and heterogeneity among rural non-farm household enterprises in Vietnam: an application of multivariate analysis

Nguyen Huu Chi*
Tran Kim Thu*
Tran Thi Bich*

* The National Economic University, Vietnam

**Vietnam Economist Annual Meeting (VEAM) 2010**

**the 3rd Conference, Cua Lo, Nghe An, Vietnam, August 24-25**

# Informality, business outcomes and heterogeneity among rural non-farm household enterprises in Vietnam: an application of multivariate analysis

**Nguyen Huu Chi*, Tran Kim Thu, Tran Thi Bich**

The National Economic University

(preliminary version)

## Abstract

This paper highlights some features in terms of informality and heterogeneity of the rural non-farm household enterprises in Vietnam. Multivariate data analysis is applied to the data set of Vietnam Household Living Standard Survey in 2004. The results show that informality is largely explained by firm size, initial capital, and economic outcomes. Cluster analysis is, then, used to identify economic typology of rural non-farm household enterprises. Five distinct groups are identified from the analysis, showing big differentials from one another in terms of economic activities and outcomes, linkages to social networks, and heads of enterprise characteristics. The results are the *first-ever* empirical evidence on informality and heterogeneity of the rural non-farm household enterprises in Vietnam, which are valuable for policy responses.

* Corresponding author (huuchi.neu@gmail.com)

## 1. Introduction

Non-farm household enterprises (NFHEs) are composed of small-scale self-employed activities. Meyer (1992) considers small and micro enterprises as 'informality' because these firms tend to fall outside the sphere of influence, regulation and support of government. Studies on NFHEs often treat this sector as homogeneous subsistence holding for rural workers who are seeking to diversify economic activities and sources of income complementing to agricultural activity. The literature on the informal sector tends to treat only the urban part of the sector, considering the informality solely a phenomenon in urban areas. There have been few analyses and empirical evidence on the informality and compositional structure in terms of economic characteristics of rural non-farm household enterprises (RNFHEs). Moreover, most of the existing studies provide evidence on RNFHEs in countries in Africa (Miles and Norcliffe, 1984; Vijverberg1988, 1990; Livingstone, 1991).

The shortage of empirical analysis on the informality and economic structure of RNFHEs may lead to a misunderstanding of the NFHEs sector. Informality is often considered as the problem of urban management. Economic policies are, therefore, less effective because they are designed especially for NFHEs in urban areas and then applied for the whole sector, without taking into account the specificity of different types of enterprises in the sector.

In this paper we address the above issues by examining the economic typology and investigating key factors explaining the heterogeneity of RNFHEs for the case of Vietnam. A multiple correspondence analysis (MCA) is applied to analyse a data set of the Vietnam Household Living Standard Survey (VHLSS) in 2004, which is considered as the unique round to provide detailed information on various aspects of non-farm business activities of households to date. The analysis allows to identify key factors explaining heterogeneity among RNFHEs. Cluster analysis is then applied to generate 5 groups of RNFHEs that differ from one another. Automatic statistical description of each cluster is generated from this analysis allows figuring out the distinct characteristics of each type of RNFHEs.

The remainder of the paper is organised as follows. Section 2 presents brief literature reviews on the informality, firms' performance and sources of heterogeneity among RNFHEs in Vietnam. Data source and empirical strategies with the emphasis on multivariate data analysis are briefly described in Section 3. Empirical results are presented in Section 4 and Section 5 concludes.

## 2. Informality and heterogeneity among rural non-farm household enterprises

In the literature RNFHEs are usually classified by different sub-groups of activities. This sector is often highlighted as comprising of a highly diverse range of activities such as manufacturing, services, etc. Other economic characteristics of enterprises have also been relied on classifying the activities, which show that the RNFHE sector is not homogenous. As regards the formality of enterprises, the vast majority of RNFHEs usually considered as informal since they have common characteristics with the urban informal sector enterprises: small-scale that are not

formally registered; unregistration of workers to social security; blurred seperation between the household and business activity, etc. For this reason there has been a viewpoint that considers splitting the whole NFHE sector into two sub-categories, urban NFHEs which constitue substantial part of the informal sector in urban area, and RNFHEs (see for instance Miles and Norcliffe, 1984). Vijverberg (1988 and 1990) based on the data obtained from the Cote d'Ivoire Living Standard Survey conducted in 1985 to provide some statistical evidence showing that rural non-farm enterprises have many characteristics which are usually found in urban informal sector enterprises, and to test for categorical identity between non-farm self-employment and the informal sector.

Heterogeneity among small-scale enterprise can be reflected in various aspects related to firms' characteristics and derived from different sources. The variety of enterprises is often revealed by firm size, the degree of formality, the motivation of entry and firm's dynamics etc. A number of studies on the heterogeneity of informal sector or small-scale household enterprises have been originated from Lucas' (1978) model of the size distribution of enterprises. He argued that the vast range of firm size among established firms can be derived from the difference in cost structures and entrepreneurial ability which result in difference in proficiency among enterprises. The more proficient entrepreneurs the larger the enterprise and the higher the business outcomes they have. Also, relying on this idea, Jovanovic (1982) presents a model interpreting heterogeneity in the entrepreneurs' dynamics. In this model enterprise's true costs of production is an adjusted cost function with a multiplier which is a function of entrepreneurial ability or advantageous location. Entrepreneurs have only vague estimate of true cost structure at the time of start up since their precise entrepreneurial ability is initially unknown. More precise estimate of enterprise's cost structure can be obtained based on actual level of profit when entrepreneurs actually operate a business. If entrepreneurs realize profits above their expected level they will revise downward their estimates of costs, and expand their business to long run size. Those who realize that their business far less profitable than expected will more likely to go out of business. Applied for the case of rural NFHEs, this model is explainable for the observed distribution of enterprises by size and formality that reflect the underlying distributions of factors contributing to entrepreneurial ability and location of enterprise.

Regarding the size and dynamics of enterprises, Cunningham and Maloney (1999) recognized that Jovanovic's model has two implications. First, the observed distribution of enterprises revealed heterogeneity when there co-exist large enterprises with high capital intensive and business outcomes with very small ones. Second, heterogeneity within small enterprises can also be realized from the co-existence of long-standing enterprises at long run size, newly established enterprises that find themselves profitable and desire to expand and newly established enterprises that are about to fail. Long lived enterprises are more likely to be found at long run size since they are more efficient than others. In analysing the dynamics of NFHEs in Vietnam, Vijveberg and Houghton (2004) have shown that older and larger firms are more likely to survive and expand. The variance in entrepreneurial abilities and preferences and the noisy

process of discovering them should be the factors explaining for the variety of enteprises in terms of size and entrepreneurial dynamics.

Turning to the distribution of formality among small-scale enterprises, Leveson and Maloney (1988) have also based on Jovanovic's model to interprete the interrelationships between size, growth and formality that are considered as sources of heterogeneity among enterprises. As indicated in this study, there is heterogeneity in the degree of formality reflected by firms' particiaption in societal institutions. Econometric results obtained in this study show positive effects of firm age and size to the probability of participation. These evidence supports their hypothesis stated that small, inefficient and young firms are disproportionately infomal and that the mode of operation and formality are jointly determined.

## 3. Data and variables

The data used in the analysis are drawn from the Vietnam Household Living Standard Survey (VHLSS) undertaken by the General Statistics Office in 2004. This is a nationally representative survey delicated to monitoring systematically the living standard of Vietnam's society and at the same time, to exercise the monitoring and assessment of the implementation of the comprehensive Poverty Alleviation and Growth strategy. This survey is based on a "classical" three stage stratified random sample (Primary sample unit: communes/wards; Secondary sample units: census enumeration areas or villages; Tertiary sample unit: households). Before 2002, this type of household survey has also been undertaken in 1993 and 1998, but known as the name Vietnam Living Standard Survey (VLSS). Since 2002 VHLSSs have been undertaken one round each two years with the questionnaire containing compatible core modules and additional modules on some specific topics for certain round.

The round of VHLSS in 2004 was designed to collect detailed information on non-farm business activities of households by incorporating an additional module on this topic. Based on the database of this survey we can extract a dataset which consist of 2,018[1] RNFHEs. Information on RNHFEs is obtained from the section 4c and 10 of the questionnaire. Section 4c is one of the core modules that provides information on few main characteristics such as branch of activity, registration of business, and the operation of RNFHEs which allow to construct economic

[1] The dataset is prepared by merging data files of different sections. In order to obtain information about charactetistics of the head of NFHEs we have merged enterprise-level dataset (sections 4c and 10) with individual-level dataset (sections 1, 2 and 3). Since the code of the household member who is the head of NFHE is not available neither in the section 4c nor in the section 10, the mergence of dataset between individual and enterprise levels was based on household identifier and industry code. Using industry code as the basis for merging can results in ambiguity for the cases that there are two or more NFHEs in a household with the same industry code (duplicate cases) so we decide to drop these duplicated NFHEs. By matching the two sections 4c and 10 we obtain 4,377 matched NFHEs. Among them there are 4,038 cases which are not duplicates in terms of industry code (i.e. more than one NFHE in the household that have the same industry code) and 339 duplicate cases. Regarding matched dataset for section 1, 2 and 3, we retain 2,356 individuals who are non-farm self-employed workers in the first job and 1,110 non-farm self-employers in the second job. All the cases involving households that have two or more individuals with the same industry code in their employment were dropped out. In the final merged dataset we obtain 3,064 NFHEs with information on characteristics of both enterprise and entrepreneur, among them 2,018 located in rural area.

variables measuring business outcomes. Section 10 is additional module providing more detailed information on various aspects of non-farm household business activities.

Using information drawn from this section we can construct some important variables related to the major objectives that our study is aiming at, i. e. informality, business outcomes of the RNFHEs and other additional variables related to specific topics such as social capital, institutional issues. In Vietnam, the *informal sector* is defined as all private unincorporated enterprises that produce at least some of their goods and services for sale or barter, are not registered (no business licence) and are engaged in non-agricultural activities[2]. Information on business registration of non-farm household businesses is available in VHLSS questionnaire, so that it can be used for distinguishing formal/informal RNFHEs. It is suggested that the firm size can also be used as an criterion for defining the informal sector. As the informal sector definition based upon firm size has been frequently used in the literature (Maloney, 1999; Gong, van Soest and Villagomez, 2004), we also profit the most available information in the dataset to define informal RNFHEs. This means that we report our results on informality of RNFHEs with two informal sector definitions. The fisrt one employs the criterion of business registration as in the official definition in Vietnam. The second one is compatible to the definition used in studies on the informal sector in Mexico in that it is based upon firm size criterion, i. e, RNFHEs which have less than 6 workers are informal.

Finally, the dataset used in this study includes four groups of variables as the followings:

*Characteristics of the entrepreneur (head of the RNFHE), including:* age; sex; years of schooling; tenure in the working field; binary variables capturing types of job-learning that the head of RNFHE has experienced (trained by tutor, formally trained, trained in paid job).

*Characteristics of the RNFHE:* years in operation (the number of years the RNFHE has been in business); business location; capital-labor ratio (K/L); size of entreprise (total number of workers in the RNFHE); binary indicating the RNFHE employed wage workers or not; the number of operating days during last 12 months; idustrial branch of activity.

*Variables of business outcomes of the RNFHE:* this group includes economic variables (production, value added, mixed incomes) constructed from the data obtained in the section 4. Production of RNFHE over period of 12 months is calculated from toatal amount of monthly turnover minus the amount of auto consumption. Value added of each RNFHE is estimated by subtracting the amount of intermediate comsumption from monthly production. Information used for the calculation of intemediate consumption is provided in section 4C2, with detail monthly amount of expenditure on 12 items of input. Mixed income of each RNFHE is then calculated by subtracting from value added the amount of mass salary and tax paid by the RNFHE.

*Entrepreneurial dynamics and difficulties faced in business:* this group can be devided into three classes of variables related to the way of trading products of the RNFHE, social network and

[2] See Cling et al. (2010) for more details

difficulties faced by the RNFHE. The sub-set of variables concerning way of trading products is mainly composed of three binary variables indicating the RNFHE has sold or not its products on each type of market (within province, other provinces and international one). The sub-group of variables considered as proxies for social networks of the RNFHE includes binary variables indicating: whether the head of RNHFE has relationship with other RNFHEs doing the same activity owned by his relatives or friends; whether initial capital of the business was financed by loans from family member, relatives or friends; and whether the RNFHE acquired information for pricing product from other traders or firm operating in the same business.

## 4. Theoretical framework

### *Multiple Correspondent Analysis*

In the first step of exploring main patterns among RNFHEs, we apply data reduction techniques. Based on the types and setting of variables used in analysis, two alternative exploratory data analysis can be used, i. e. Principle Component Analysis (PCA) and Multiple Correspondent Analysis (MCA). PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The components are created to account for maximal variation among the original variables, i.e. the first principal component accounts for as much of the variability in the initial variables as possible, and each succeeding component accounts for as much of the remaining variability as possible. MCA is a variant of PCA aimed primarily at categorical data (Benzécri, 1979; Greennacre, 1984, 2006; Lebart, Morineau and Piron, 2006).

In this paper, MCA is applied for the extended set of variables concerning characteristics of RNFHEs in the form of categorical data such as types of business, location, formality, obstacles faced by the RNFHE. Other quantitative variables such as life of enterprise, capital to labor ratio, value added per worker, etc. can also be coded into categories and included in MCA.

The original data matrix (X) to be analyzed in this case is an [RNFHEs × Categorical Variables] matrix, where the rows represent RNFHEs and the columns display categories of the original variables. MCA can be defined in many equivalent ways, but the two principle ones are MCA based on the decomposition of indicator matrix or Burt matrix. Indicator matrix is the respondents-by-categories matrix of dummy variables. Each categorical variable generates as many zero/one dummy variable as response categories and these variables form the column of the indicator matrix while the rows represent RNHHEs. Burt matrix is the matrix of associations among variables involved all the cross-tables of variables chosen in the analysis with themselves. This forms a data structure is denoted **C**, which is a square supermatrix of cross-tables and related to indicator matrix by the formula $\mathbf{C} = K^T K$.

Assuming that we have in the original data Q categorical variables, the indicator matrix K can be presented as $K = [K_1, K_2, \dots, K_Q]$. Let denote the number of categories for the q*th* categorical

variable $J_q$ , and the total number of categories $J = \sum J_q$. Similar to PCA, MCA is based on decompositions of centered and normalized matrices, using either the eigenvalue-eigenvector decomposition of a squared symmetric matrix or the singular-value decomposition (SVD) of a rectangular matrix (Greennacre, 2006). For the case of performing CA of the indicator matrix, we take SVD as the following equation:

$$\sqrt{n}\left(\frac{\mathrm{K}}{\mathrm{Qn}} - \frac{1}{\mathrm{n}}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Gamma}^{2}\mathbf{V}^{\mathrm{T}} \qquad (4.1)$$

where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, and (1/n)$\mathbf{1}$ is the vector of row masses; $\mathbf{1}^{\mathrm{T}}\mathbf{D}$ is the vector of column masses of the indicator matrix **K**.

For the case of decomposition taken for centered Burt matrix as in the following equation:

$$\mathbf{D}^{-\frac{1}{2}}\left(\frac{\mathrm{C}}{\mathrm{Q}^{2}\mathrm{n}} - \mathbf{D}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-\frac{1}{2}} = \mathbf{V}\boldsymbol{\Gamma}^{2}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \qquad (4.2)$$

where, $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$; $\mathbf{D}$=1/2diag$(\mathbf{D_1}, \mathbf{D_2}, \ldots, \mathbf{D_Q})$ is the marginal relative frequency matrix for the column of **C**.

In the above equations, $\boldsymbol{\Gamma}$ is the diagonal matrix of singular values, and $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}^2$ is the matrix of eigenvalues**.** The maximum value of the average squared correlation is the square of the first sigular value in $\boldsymbol{\Gamma}$ or (eigenvalue - $\lambda_1$ ) in the centered analysis of **C.** In MCA the eigenvalues are also called principal inertias. The percentages of inertia represent percentages of explained variance in the data matrix.

No matter whether CA is performed on indicator matrix or Burt matrix, the decomposition provides the same factor scores for the rows (RNFHEs) and the columns (modalities of variables exhibiting the characteristics of RNFHEs) of the indicator matrix. These factor scores are scaled such that their variance is equal to their corresponding eigenvalue. The row and comlumn factor scores are obtained respectively by the following formulas : $\mathbf{F} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Gamma}^{2}$ and $\mathbf{G} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Gamma}^{2}$**.**

In interpreting the obtained factors, we base on the results reported on the contribution and squared cosin between each row or each column and each factor. The contributions of row *i* or column *j* to factor *s* help defining the importance of RNFHE *i* or modality *j* to factor *s*. These contributions can be respectively presented as: $t_{i,s} = \frac{r_i f_{i,s}^2}{\lambda_s}$ and $t_{j,s} = \frac{c_j g_{j,s}^2}{\lambda_s}$ , where $r_i$ and $c_j$ are the total of row *i* and column *j* respectively. The squared cosin between row *i* and factor *s* and column *j* and factor *s* are obtained by the formulas: $cos_{i,s}^2 = \frac{f_{i,s}^2}{d_{r,i}^2}$ and $cos_{j,s}^2 = \frac{g_{j,s}^2}{d_{c,j}^2}$ , where $d_{r,i}^2$ and $d_{c,j}^2$ are the squared distance from the row *i* and column *j* to the central gravity of points in the respective spaces of individuals (RNFHEs) or modalities. The squared cosin help assessing the quality of representation of modalities or individuals (RNFHEs) on factor *s*. The closer to 1 this indicator is, the better relative contribution the factor *s* has on the position of the modality (or RNFHE).

As presented above, MCA recodes the original data by creating indicator matrix in which several binary columns are used to capture data for each categorical variable. This makes the total inertia of the solution space artificially inflated. Therefore, the percentage of inertia explained by the first dimension is underestimated. To overcome this issue, we can use a correction formula developed by Benzecri (1979) as follows:

$$\lambda_s^* = \left[\frac{Q}{Q-1}\left(\lambda_s - \frac{1}{Q}\right)\right]^2 \qquad (4.3)$$

Where $\lambda_s$ is the eigenvalues obtained from the analysis of the indicator or Burt matrix; $\lambda_s^*$ is the corrected eigenvalues.

***Cluster analysis***

The objective of applying this method is to group RNFHEs into clusters described by a set of RNFHE's characteristics or categories of qualitative variables, so that the RNFHEs within a cluster are as similar to each other as possible and the RNFHEs between clusters dissimilar as much as possible. In clustering the RNFHEs, we follow a procedure which consists of five steps presented as follows:

- *Selecting variables for analysis:* As the cluster analysis is undertaken as the second phase after MCA, the variables selected are factors generated form MCA which help reducing the complexity of the dimensional space of the original variables.

- *Selecting a distance measure:* Cluster analysis is performed based on assessing similarity. In the dimensional space obtained through MCA, the similarity between units, i.e., RNFHEs is assessed by measuring the distance between units. The closer the units are, the more similarity they have. The most common measure used in cluster analysis is Euclidian distance which is based on Pythagorea's theorem to assess the shortest distance between two points. It is noted, however, that if cluster analysis is performed basing on the original categorical data then the suitable distance measure is chi-square distance[3]

- *Selecting a clustering procedure:* Clustering procedure is an algorithm that helps producing clustering solutions. These algorithms can be hierarchical, non hierarchical, or a two-step procedure. Our empirical analysis is relied on Data and Text Mining (DtmVic) version 4.0 developed by Lebart *et al.* (2008) to process clustering analysis. In this package, clustering is based on hybrid method using both hierarchical clustering and k-means (non hierarchical) technique.

When applying hierarchical clustering algorithm in classifying RNFHEs, the procedure starts with assigning each RNFHE to its own cluster and interatively joins together the two closest clusters. This means that the distances betweens RNFHEs are used as criterion of assessing the similarity of RNFHEs. The Euclidian distances calculated among RNFHEs in the dimensional space formed by factorial axis serve as input to the clustering algorithm. The joining process of

[3] See Greenacre (1984) for more details

similar clusters continues until all the RNFHEs are clustered into a single cluster. The output of this process is a hierarchical tree or dendrogram. Dendrogram is a binary tree in which the leaves are the original single element clusters and the clustering process is represented by horizontal lines.

K-means[4] clustering is a non-hierarchical method. The application of this method of clustering in classifying RNFHEs envolves a four basic operation process: *(1)* selection of initial k means for k clusters; *(2)* calculation of the dissimilarity between an RNFHE and the mean of a cluster; *(3)* allocation of the RNFHE to the cluster whose mean is nearest to the RNFHE; *(4)* re-calcualation of the mean of cluster from the RNFHE allocated to it. After this process the intra cluster dissimilarity is minimised.

- *Deciding the number of clusters*: this step is done by cutting the dendrogram at a given height. Choosing the level of the cut, and thus the number of clusters, can be facilitated by visual inspection of the dendrogram as well as the histogram of indices of level. Cutting off the tree can be facilitated by examining the histogram of the increasing level of indices. It should be cut at the height at which the histogram represents a significant step. Each bar histogram indicates that the index value of an aggregation, i.e the loss of inertia obtained from one partition of s clusters to to the partition of s - 1 clusters.

- *Interpreting the profile of clusters:* the interpretation is based on a set of variables and the associated modalities that can be best used to characterize each cluster. In defining these variables, inferential comparisons of means or percentages of RNFHEs in a cluster with the means or percentages on all RNFHEs to be classified are undertaken. Value tests are performed for determining which are the significantly correlated modalities with the cluster. Those modalities positively correlated with the analyzed cluster are representable for that cluster.

## 5. Empirical results

### 5.1. Informality among RNFHEs: global picture

***High rate of informality***

Rural NFHE sector in Vietnam is also characterized by high rate of informality. As mentioned above, in defining informal NFHEs, we refer to informal sector definitions applied in other studies in Vietnam. Table 5.1 reports the proportion of formal and informal NFHEs for each region in Vietnam based on two operational definitions: the first one defines informal NFHEs as those having business license; the second one adds one more criterion concerning the size of NFHEs in terms of the number of workers. As in the presented results, there is slight difference between the figures obtained by using the two informal sector definitions. This also reflects that there is a cohesive relation between the size and formal registration of NFHEs. On average, the informal NFHEs account for about 85% of all rural NFHEs. This result seems to be highly

[4] A variant of k-means algorithm applied for the case of clustering based on original categorical variables is k-modes described in Huang (1997)

accordant with some findings in other survey and studies on small-scale rural household enterprises. For instance, Vu (2006) found some statistical evidence indicating that the informal household business is the main business form in rural area. The case of Bac Ninh province can be a good example since it is reported in Bac Ninh DPI 2005 that 89% of rural household business were home based and did not have business license. Recent results obtained from Labor Force Survey (LFS 2007) also show that the rate informality is higher in rural and suburban areas (see Cling *et al.*, 2010). Regarding the geographical differences, it can be seen that the Northeast and Red River Delta are the two regions that have the highest rates of informality (90% and 93% respectively) of rural NFHEs.

**Table 5.1: Informality among rural NFHEs by regions**

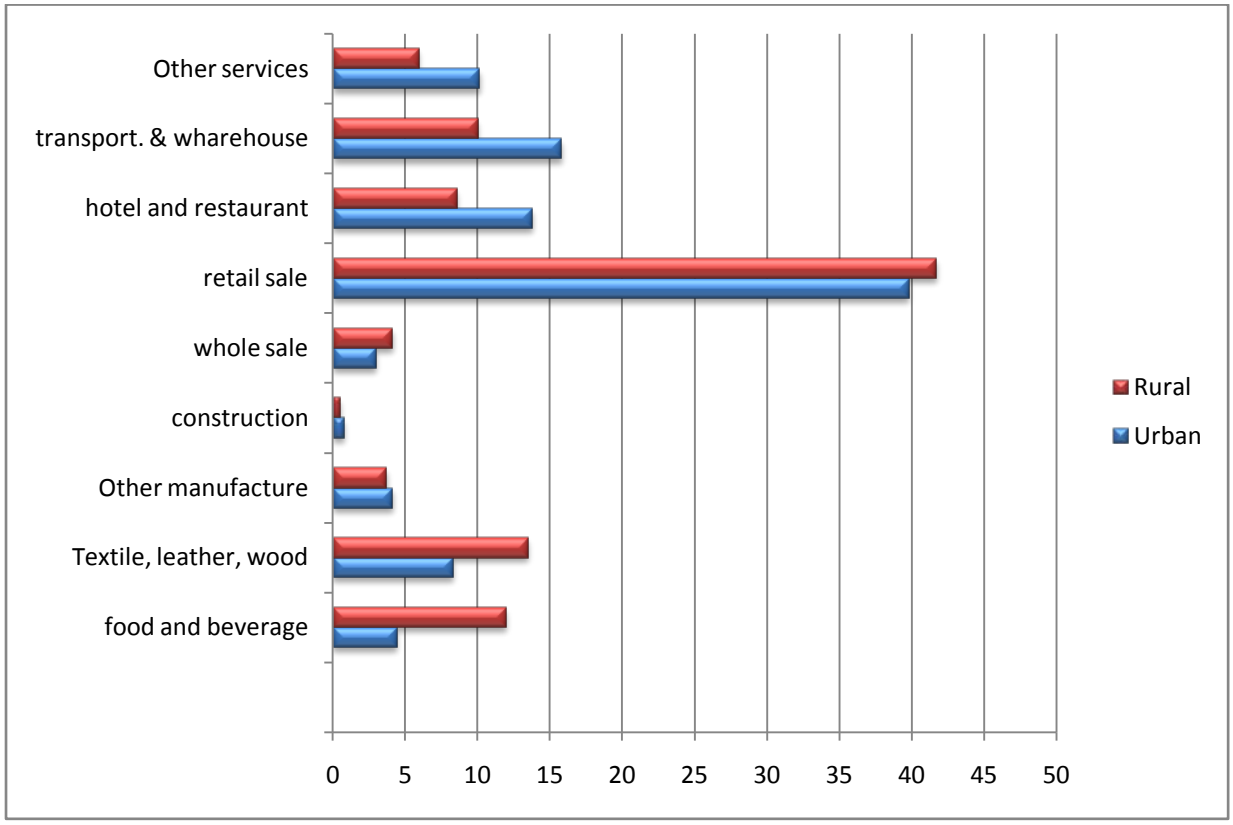
| Region | Registration definition | | Firm-size definition | | Total |
|---|---|---|---|---|---|
| | Formal | Informal | Formal | Informal | |
| Red River Delta | 9.4 | 90.6 | 10.2 | 89.8 | 100 |
| North West | 21.5 | 78.5 | 21.5 | 78.5 | 100 |
| North East | 7.0 | 93.0 | 7.0 | 93.0 | 100 |
| North Central Coast | 14.5 | 85.5 | 15.6 | 84.4 | 100 |
| Central Coast | 16.6 | 83.4 | 16.6 | 83.4 | 100 |
| Central Highlands | 21.6 | 78.4 | 22.7 | 77.3 | 100 |
| Southeast | 15.9 | 84.1 | 15.9 | 84.1 | 100 |
| Mekong River Delta | 14.3 | 85.8 | 14.7 | 85.3 | 100 |
| Total | 14.6 | 85.4 | 15.1 | 84.9 | 100 |

*Source: VHLSS 2004, authors' calculation*

***Difference of sectoral composition between rural and urban informal NFHEs***

Figure 5.2 presents sectoral composition of rural and urban informal NFHEs. It is clearly revealed the predominance of informal NFHEs engaging in retail sale business in both rural and urban areas. The difference of rural informal NFHEs from their urban counterparts can be realized for those that involve manufacturing and service activities. The proportion of informal household enterprises in manufacturing is significantly higher in rural area whereas urban informal household enterprises are found more in services, especially those involve accommodation and transportation. These differences can be explained by the specificity of each locality. The higher representation of manufacturing activities such as food and beverage processing or wood and rattan production is possibly associated with the systematic linkage between these rural NFHEs and agricultural activities.

**Figure 5.2: Structure of rural and urban informal NFHEs by fields of activity**




*Source: VHLSS 2004, authors' calculation*

### 5.2.Main factors attributed to as sources of heterogeneity among RNFHEs

In the first step of this analysis, we use a multiple correspondent analysis (MCA)[5] in order to regroup variables and reduce the number of dimensions contained in the primary dataset. This means that we obtain main dimensions based on various initial variables relating to characteristics of RNFHEs and entreprneneurs.

Two groups of variables are distinguished when taking this analysis. Firstly, active variables are composed of entrepreneurs characteristics such as age, education, and tenure as well as enterprise characteristics such as firm life, number of workers, use of wage workers, capital-labor ratio. Totally, 14 active variables with 52 associated modalities have been selected (see Table 5.3). These active variables will contribute to the performance of CA and explain to the factors derived from the analysis. Secondly, 22 supplementary variables with 74 associated modalities are also introduced in the the analysis in order to provide more informative results on the related issues to the heterogeneity of RNFHEs, especially on job-learning, social networks and institution-related diffculties. Thus, these illustrative variables can be distinguished by three sub-groups. The first one includes binary variables capturing types of job-learning that the entrepreneurs have experienced as well as other proxies for social networks (such as dummies

[5] The analysis is carried out by using the Data and Text Mining sofware (DtmVic 4.0) developped by Lebart et al. (2008).

indicating the entrepreneurs have relative or friendship with other RNFHEs producing the same products; getting finance for initial capital from relatives or friends or not; being a member of business associations or not). It is referred to in several studies on social job-learning and social networks that these are factors correlating to entrepreneurial ability to perform efficiently business activity at lower transaction cost (Zuwarimwe and Kirsten, 2007). Therefore, the visualisation of the associated modalities of these variables on the dimensional planes formed by factorial axis is expected to help providing more explanation for the variety in the characteristics and business outcomes of RNFHEs. The second sub-group of supplementary variables consists of those revealing obstacles faced by RNFHEs. Five binary variables capturing information on difficulties concerning different business resources (electricity, communication, transportation, land and finance) and seven categorical variables providing information on the degree of difficulty that the RNFHE faced in six institution-related aspects which are taxes (variable name is **pbtax**), business registration and operation registration (**pbbusire**), labor regulations (pblabore), commercial and customs regulations (**pbcomreg**), inconsistent economic policies (**pbecopol**) and corruption (**pbcorrup**). For each of these variables, the possible answers are "no hurdles" (coded as "cat_1"), "small hurdles" ("cat_2"), "noteworthy hurdles" ("cat_3"), "major hurdles" ("cat_4"), "serious hurdles" ("cat_5"). "irrelevant" and "unknown" are respectively coded as "cat_6" and "cat_7".

**Table 5.3: Active variables used in MCA**

| Variables | Description | Description of modalities |
|---|---|---|
| **Agegr** | age of the head of RNFHE | **agegr_cat_1**: <24 y.o; **agegr_cat_2**: 24-34 y.o; **agegr_cat_3**: 35-44 y.o; **agegr_cat_4**: 45-54 y.o; **agegr_cat_5**: >=55 y.o |
| **Educa** | education attainment of the head of RNFHE | **educa_cat_1**: no degree**; educa_cat_2**: primary; **educa_cat_3**: lower secondary; **educa_cat_4**: upper secondary; **educa_cat_5**: university & others |
| **tenuregr** | tenure of the head of RNFHE in the current job | **tenuregr_cat_1**: <3 years; **tenuregr_cat_2**: 3 - 5 years; **tenuregr_cat_3**: 6 - 10 years; **tenuregr_cat_4**: >=11 years |
| **informal** | informality of RNFHE | **informal_cat_0**: formal RNFHE; **informal_cat_1**: informal RNFHE |
| **Lifehbgr** | RNFHE's life | **lifehbgr_cat_1**: <3 years; **lifehbgr_cat_2**: 3 - 5 years; **lifehbgr_cat_3**: 6 - 10 years; **lifehbgr_cat_4**: >=11 years |
| **Qinicapi** | quintiles of initial capital of RNFHEs | **qinicapi_cat_1**: 1st quintile; **qinicapi_cat_2**: 2nd quintile; **qinicapi_cat_3**: 3st quintile ; **qinicapi_cat_4**: 4th quintile |
| **sizehbgr** | categories of RNFHEs by the number of workers | **sizehbgr_cat_1:** own account; **sizehbgr_cat_2:** 2 workers; **sizehbgr_cat_3:** 3 - 4 workers; **sizehbgr_cat_4:** >=5 workers |
| **usepaidw** | RNFHE has paid workers | **usepaidw_cat_0:** RNFHE has not paid worker; **usepaidw_cat_1:** RNFHE has paid worker; |
| **Soldothe** | RNFHE sell product in other provinces | **soldothe_cat_0:** does not sell in other prov. **soldothe_cat_1:** sell product in other prov. |
| **branch3** | branch of activity | **branch3_cat_1:** manufaturing; **branch3_cat_2:** commerce; **branch3_cat_3:** services |
| **Busiloca** | business location | **busiloca_cat_1:** home; **busiloca_cat_2:** unfixed place; **busiloca_cat_3:** industrial park/commercial centre; **busiloca_cat_4:** professional shop or fixed place in the market. |

| | | |
|---|---|---|
| **qmprod** | quintiles of production of RNFHEs | **qmprod_cat_1**: 1st quintile; **qmprod_cat_2**: 2nd quintile; **qmprod_cat_3**: 3st quintile ; **qmprod_cat_4**: 4th quintile |
| **qmvalpel** | quintiles of value value added per worker | **qmvalpel_cat_1**: 1st quintile; **qmvalpel_cat_2**: 2nd quintile; **qmvalpel_cat_3**: 3st quintile ; **qmvalpel_cat_4**: 4th quintile |
| **qmearnin** | quintiles of earnings of RNFHEs | **qmearnin_cat_1**: 1st quintile; **qmearnin_cat_2**: 2nd quintile; **qmearnin_cat_3**: 3st quintile ; **qmearnin_cat_4**: 4th quintile |

A cleaning step is undertaken in MCA procedure resulted in the elimination of some modalities with too small weight so that there remains 14 active variables associated with 49 modalities. The decomposition of Burt matrix results in a diagonal matrix of eigenvalues (principal inertias) and associated matrix of standard coordinates. The principle inertias, the pecentages and acumulative percentage of these inertias and histogram are presented in Figure 5.4.

**Figure 5.4: histogram of the 32 first eigenvalues (principal inertias)**

| number | Eigen value | percent. | cumulat. percent. | |
|---|---|---|---|---|
| 1 | .2450 | 10.27 | 10.27 | ********** |
| 2 | .1827 | 7.66 | 17.93 | ******* |
| 3 | .1598 | 6.70 | 24.64 | ****** |
| 4 | .1397 | 5.86 | 30.49 | ***** |
| 5 | .1215 | 5.09 | 35.59 | ***** |
| 6 | .1130 | 4.74 | 40.33 | ***** |
| 7 | .1064 | 4.46 | 44.79 | ***** |
| 8 | .0979 | 4.10 | 48.89 | **** |
| 9 | .0926 | 3.88 | 52.77 | **** |
| 10 | .0844 | 3.54 | 56.31 | **** |
| 11 | .0802 | 3.36 | 59.67 | **** |
| 12 | .0771 | 3.23 | 62.91 | **** |
| 13 | .0768 | 3.22 | 66.13 | **** |
| 14 | .0727 | 3.05 | 69.18 | *** |
| 15 | .0705 | 2.96 | 72.13 | *** |
| 16 | .0683 | 2.86 | 75.00 | *** |
| 17 | .0640 | 2.68 | 77.68 | *** |
| 18 | .0617 | 2.59 | 80.26 | *** |
| 19 | .0609 | 2.55 | 82.82 | *** |
| 20 | .0565 | 2.37 | 85.19 | *** |
| 21 | .0526 | 2.21 | 87.39 | ** |
| 22 | .0519 | 2.18 | 89.57 | ** |
| 23 | .0448 | 1.88 | 91.45 | ** |
| 24 | .0421 | 1.76 | 93.21 | ** |
| 25 | .0385 | 1.61 | 94.82 | ** |
| 26 | .0367 | 1.54 | 96.36 | ** |
| 27 | .0293 | 1.23 | 97.59 | * |
| 28 | .0192 | .81 | 98.40 | * |
| 29 | .0187 | .79 | 99.19 | * |
| 30 | .0149 | .62 | 99.81 | * |
| 31 | .0045 | .19 | 100.00 | * |

*Source: VHLSS 2004, authors' calculation*

The total variance (inertias in the new space) in the original data matrix can be calculated by taking the trace of the diagonal matrix of eigenvalues which equals the sum of all the eigenvalues and is equal to 2.385. Following the method developed by Benzecri (1979) to address the problem of the inflation of the total inertias due to the fitting of the diagonal sub-matrices of the Burt matrix, we canculate modified inertias by using the formula 4.3. We selected only those eigenvalues $\lambda_p > 1/Q$ which correspond to the first 27 factors for re-calculating the inertias. The results are presented in Table 5.5.

**Table 5.5: The modified inertias**

| Factor | Eigenvalue $\lambda_p$ | $\lambda_p - \frac{1}{Q}$ | $\lambda_p^*$ | $\lambda_p^*/\Sigma\lambda_p^*$ | Cumulate percentage |
|---|---|---|---|---|---|
| 1 | 0.2450 | 0.2242 | 0.0524 | 28.1 | **28.1** |
| 2 | 0.1827 | 0.1619 | 0.0273 | 14.7 | **42.8** |
| 3 | 0.1598 | 0.1390 | 0.0201 | 10.8 | **53.6** |
| 4 | 0.1397 | 0.1189 | 0.0147 | 7.9 | **61.5** |
| 5 | 0.1215 | 0.1007 | 0.0106 | 5.7 | **67.2** |
| 6 | 0.1130 | 0.0922 | 0.0089 | 4.8 | 72.0 |
| 7 | 0.1064 | 0.0856 | 0.0076 | 4.1 | 76.1 |
| 8 | 0.0979 | 0.0771 | 0.0062 | 3.3 | 79.4 |
| 9 | 0.0926 | 0.0718 | 0.0054 | 2.9 | 82.3 |
| 10 | 0.0844 | 0.0636 | 0.0042 | 2.3 | 84.5 |
| 11 | 0.0802 | 0.0594 | 0.0037 | 2.0 | 86.5 |
| 12 | 0.0771 | 0.0563 | 0.0033 | 1.8 | 88.3 |
| 13 | 0.0768 | 0.0560 | 0.0033 | 1.8 | 90.1 |
| 14 | 0.0727 | 0.0519 | 0.0028 | 1.5 | 91.6 |
| 15 | 0.0705 | 0.0497 | 0.0026 | 1.4 | 92.9 |
| 16 | 0.0683 | 0.0475 | 0.0023 | 1.3 | 94.2 |
| 17 | 0.0640 | 0.0432 | 0.0019 | 1.0 | 95.2 |
| 18 | 0.0617 | 0.0409 | 0.0017 | 0.9 | 96.2 |
| 19 | 0.0609 | 0.0401 | 0.0017 | 0.9 | 97.1 |
| 20 | 0.0565 | 0.0357 | 0.0013 | 0.7 | 97.8 |
| 21 | 0.0526 | 0.0318 | 0.0011 | 0.6 | 98.4 |
| 22 | 0.0519 | 0.0311 | 0.0010 | 0.5 | 98.9 |
| 23 | 0.0448 | 0.0240 | 0.0006 | 0.3 | 99.2 |
| 24 | 0.0421 | 0.0213 | 0.0005 | 0.3 | 99.5 |
| 25 | 0.0385 | 0.0177 | 0.0003 | 0.2 | 99.6 |
| 26 | 0.0367 | 0.0159 | 0.0003 | 0.1 | 99.8 |
| 27 | 0.0293 | 0.0085 | 0.0001 | 0.0 | 99.8 |
| | **2.3849** | **1.7391** | **0.1862** | **100** | |

*Source: VHLSS 2004, authors' calculation*

As can be seen from Table 5.4 and 5.5, the modified eigenvalues make it clearer when comparing the contribution of each factor to total inertias. The first five factors that initially explained 10.3%, 7.7%, 6.7%, 5.8% and 5.1% respectively, with a total of only 35.6% of the total variance, now explain 28.1%, 14.7%, 10.8%, 7.9% and 5.7% with a total of 67.2% of the total variance. Hereafter we consider the factorial plane formed by each paire among the most variance explained factors. In interpreting the dimensional space, we based on the results presented in both about coordinates, contributions and cosin squared of the modalities on the first five factors[6], and factorial planes (Figure 5.8 and 2.4)

The ***fisrt factorial axis*** *"Informality and related characteristics"* shows significant loadings of the categories of variables related to formality/ informality, size of RNFHEs in terms of the

[6] The DtmVic 4.0 software provides numerical results for only 5 most variance-explained factors

number of workers and business outcomes of RNFHE in terms of value added or earnings (see Table 5.6 for absolute and relative contributions). It clearly differentiates between the two extreme situations with, at the top (see also Figure 5.8), formal RNFHEs which have 2 workers or more with high initial capital, and at the bottom are located informal RNFHEs of mostly only 1 worker with initial capital level belong to the lowest quintile group. In other words, the coordinates on this axis can be taken as a discriminant score with the rule that indicates positive values for "formal RNFHE" and negative values for "informal RNFHE". Besides explaining formality/informality, this factor reveals also, to certain extent, the business comportement and outcomes of RNFHEs. A sharp distinction can be seen for the two polarities of formal and informal RNFHEs is that the former mention to those who sell product not only in the local market but also in other provinces and obtain high productivity as well as earnings (belong to the highest quintiles), and the latter invloves less dynamic RNFHEs with the lowest productivity and earnings level and whose products are sold only in the local market.

**Table 5.6: Absolute contributions and squared cosines of variables and associated modalities of responses that contribute to the first factorial axis**

| Variables/modalities | Sign of coordinate | Absolute contributions | Squared cosines |
|---|---|---|---|
| Informal (yes) | - | 0.6 | 0.12 |
| Informal (no) | + | 3.2 | 0.12 |
| Use wage workers (no) | - | 0.6 | 0.27 |
| Use wage workers (yes) | + | 8.0 | 0.27 |
| Size of NFHE (1 worker) | - | 0.8 | 0.15 |
| Size of NFHE (2 workers) | + | 1.9 | 0.07 |
| Size of NFHE (3-4 workers) | + | 2.7 | 0.09 |
| Sell products in other provinces (no) | - | 0.2 | 0.08 |
| Sell products in other provinces (yes) | + | 2.3 | 0.08 |
| Initial capital ($1^{st}$ quintile) | - | 2.2 | 0.10 |
| Initial capital ($4^{rd}$ quintile) | + | 4.7 | 0.19 |
| Value added per labor ($1^{st}$ quintile) | - | 7.0 | 0.35 |
| Value added per labor ($4^{rd}$ quintile) | + | 10.7 | 0.44 |
| Earnings ($1^{st}$ quintile) | - | 8.2 | 0.41 |
| Earnings ($4^{rd}$ quintile) | + | 14.2 | 0.59 |

*Source: VHLSS 2004, authors' calculation*

The ***second factorial axis*** explains *"production scale of RNFHEs, sector of activity and business location"*. It opposses two types of RNFHEs . On the positive side of this factorial axis appear the modalities characterizing service RNFHEs that operate in fixed professional business location, whereas we find manufacturing RNFHEs of own-account workers working at home locate on the negative side of the axis (see Table 5.7). Modalities for quintiles of RNFHEs' production are dispersively positioned on both positive and negative sides of this axis. However, it seems to provide a clear contrast only the RNFHEs having production scale in the lowest group (the $1^{st}$ quantile) from those belong to the second and third quintiles, but not from the forth one.

**Table 5.7: Absolute contributions and squared cosines of variables and associated modalities of responses that contribute to the second factorial axis**

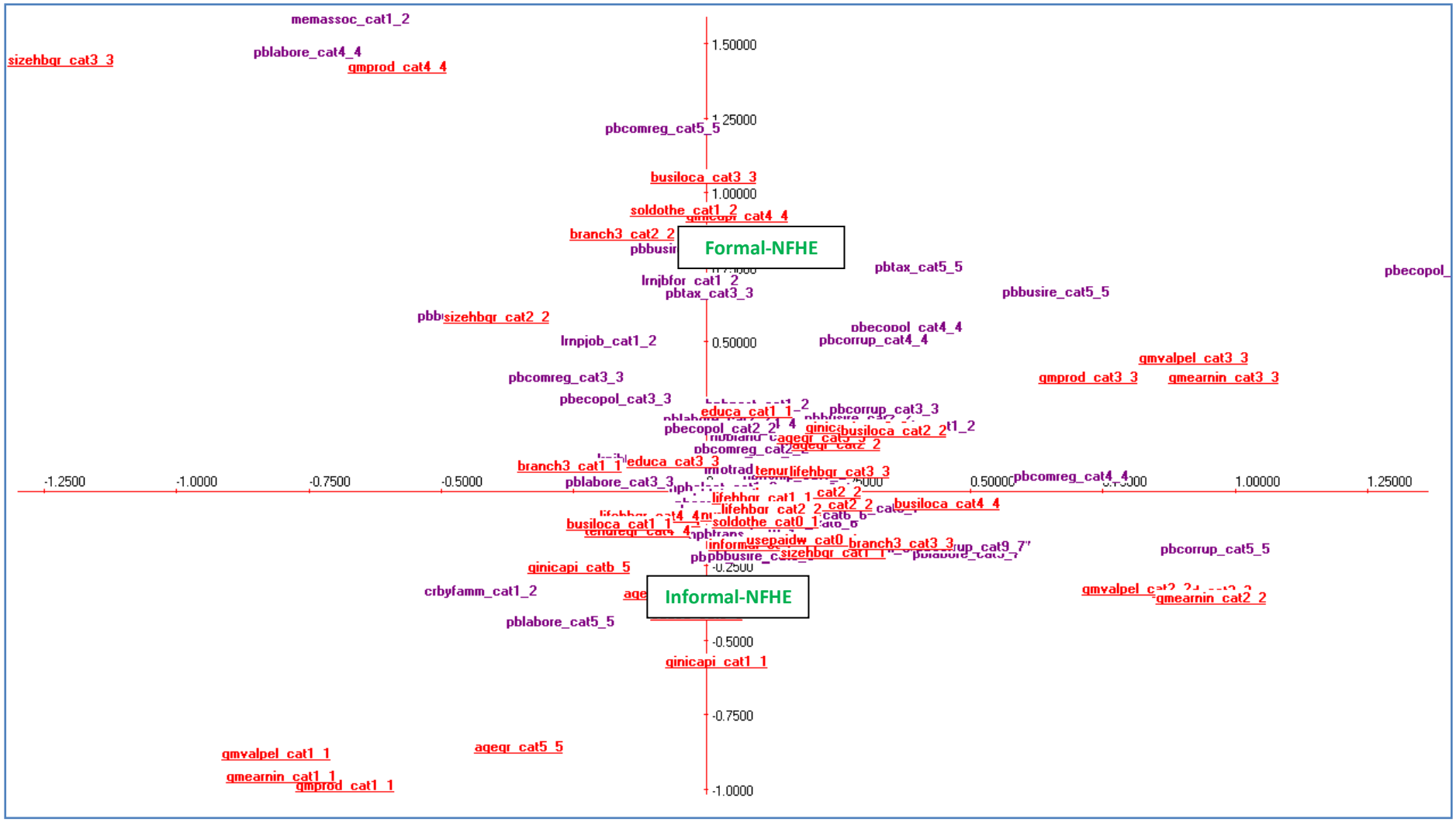
| Variables/modalities | Sign of coordinate | Absolute contributions | Squared cosines |
|---|---|---|---|
| Branch_manufacturing | - | 2.1 | 0.09 |
| Branch_services | + | 1.7 | 0.08 |
| Business location (at home) | - | 1.6 | 0.08 |
| Business location (fixed place) | + | 1.4 | 0.05 |
| Production (1st quintile) | - | 7.7 | 0.26 |
| Production (2nd quintile) | + | 8.2 | 0.27 |
| Production (3rd quintile) | + | 3.6 | 0.11 |
| Production (4th quintile) | - | 3.8 | 0.11 |

*Source: VHLSS 2004, authors' calculation*

Figure 5.8 present the factorial plane formed by the fisrt two factors containing the largest variance among original variables. This plane best explains (42.8% of the total variance) the space of positions that allow best describe the NFHB population. On this factorial plane we observe together the location of modalities of both active variables which contributed to the formation of factorial axis and of supplementary variables. The analysis of the supplementary variables provides additional information about the polarity of each factorial axis.

As for the first axis, on the positive side, where formal RNFHEs are, we found closely located some very informative modalities such as **lrnjbfor_cat1** (learned job through formal training), **memassoc_cat1** (the entrepreneur was a member of professional associations), **pbtax_cat5** (have serious hurdles concerning taxes), **pbcomreg_cat5** (have serious hurdles concerning commercial and customs regulations). On the other side, we found that informal RNFHEs of own account workers are associated with modalities such as **lrnjbfor_cat0** (did not have any formal training), **lrnjbtu_cat0** (did not have any tutorial training), and various modalities coded either cat1 (no hurdles), cat2 (small hurdles) or cat6 cat7 (irrelevant) of variables concerning institutional obstacles such as taxes, business registration, labor regulation, corruption, etc. These results suggest two implications on the informality of RNFHEs. First, there was a positive correlation between formality of RNFHEs and the mode of job learning that the entrepreneur acquired. Those who had certain level of training for current job were more likely to formally register their business. Second, the correlation between formality and the incidence of facing with institutional obstacles implies barriers to formalization of RNFHEs. This is consistent with the results obtained from a survey recently and speciffically undertaken to study the informal sector in Hanoi and Ho Chi Minh city which was more concentrated on urban centers (see Cling et al., 2010). The surveyed results shown that formal household businesses in these cities were barely more affected by problems with public authorities, mainly with the police and tax administration.

**Figure 5.8: plane of factors 1 and 2**



Now we turn to the ***third factorial axis*** which explains *"life of RNFHEs and seniority of entrepreneurs"*. It characterize a clear contrast between two poles, on one side (the positive semi axis), RNFHEs that have longevity of less than 5 years and operated by young entrepreneurs (particularly those aged below 24 years old) with low level of seniority, on the other side, those that have been entered in operation for 11 years or more owned by senior self-employers aged 55 years old and more (see Table 5.9). Regarding the illustrative variables, it can be seen in the dimensional space formed by the first and the third factorial axis (see figure 5.12) that located along the third factorial axis polarized modalities of some variables concerning institution-related problems faced by RNFHEs such as **pbcorrup_cat5** (having serious hurdles concerning corruption) and **pbcomreg_cat4** (having major hurdles concerning commercial and customs regulations), both positioned on the negative semi axis. This means that, to some extent, the third factor is associated with the possibility of RNFHEs' facing these obstacles.

**Table 5.9: Absolute contributions and squared cosines of variables and associated modalities of responses that contribute to the third factorial axis**

| Variables/modalities | Sign of coordinate | Absolute contributions | Squared cosines |
|---|---|---|---|
| Age of the head of NFHE (<24 years old) | - | 8.0 | 0.19 |
| Age of the head of NFHE (>=55 years old) | + | 4.3 | 0.10 |
| Tenure of the head of NFHE (<3 years) | - | 12.6 | 0.35 |
| Tenure of the head of NFHE (3-5 years) | - | 2.9 | 0.9 |
| Tenure of the head of NFHE (6-10 years) | + | 3.1 | 0.9 |
| Tenure of the head of NFHE (>=11 years) | + | 15.4 | 0.41 |
| Life of NFHE (<3 years) | - | 8.0 | 0.19 |
| Life of NFHE (3-5 years) | - | 7.0 | 0.22 |
| Life of NFHE (>=11 years) | + | 16.7 | 0.49 |

*Source: VHLSS 2004, authors' calculation*

Together with the first factor, the forth factor explains business outcomes disparity among the two quintiles located in the middle position on the distributions of NFHEs by value added and earnings. Business outcomes variables such as value added per labor and earning do not entirely contribute to the first factor. They participate in characterizing the fifth axis with the opposition of the modalities of the two middle quantiles of either value added per labor or earnings of NFHEs.

**Table 5.10: Absolute contributions and squared cosines of variables and associated modalities of responses that contribute to the forth factorial axis**

| Variables/modalities | Sign of coordinate | Absolute contributions | Squared cosines |
|---|---|---|---|
| Value added per labor ($2^{nd}$ quintile) | - | 15.1 | 0.38 |
| Value added per labor ($3^{rd}$ quintile) | + | 9.9 | 0.24 |
| Earnings ($3^{rd}$ quintile) | + | 15.4 | 0.39 |
| Earnings ($2^{nd}$ quintile) | - | 15.0 | 0.35 |

*Source: VHLSS 2004, authors' calculation*

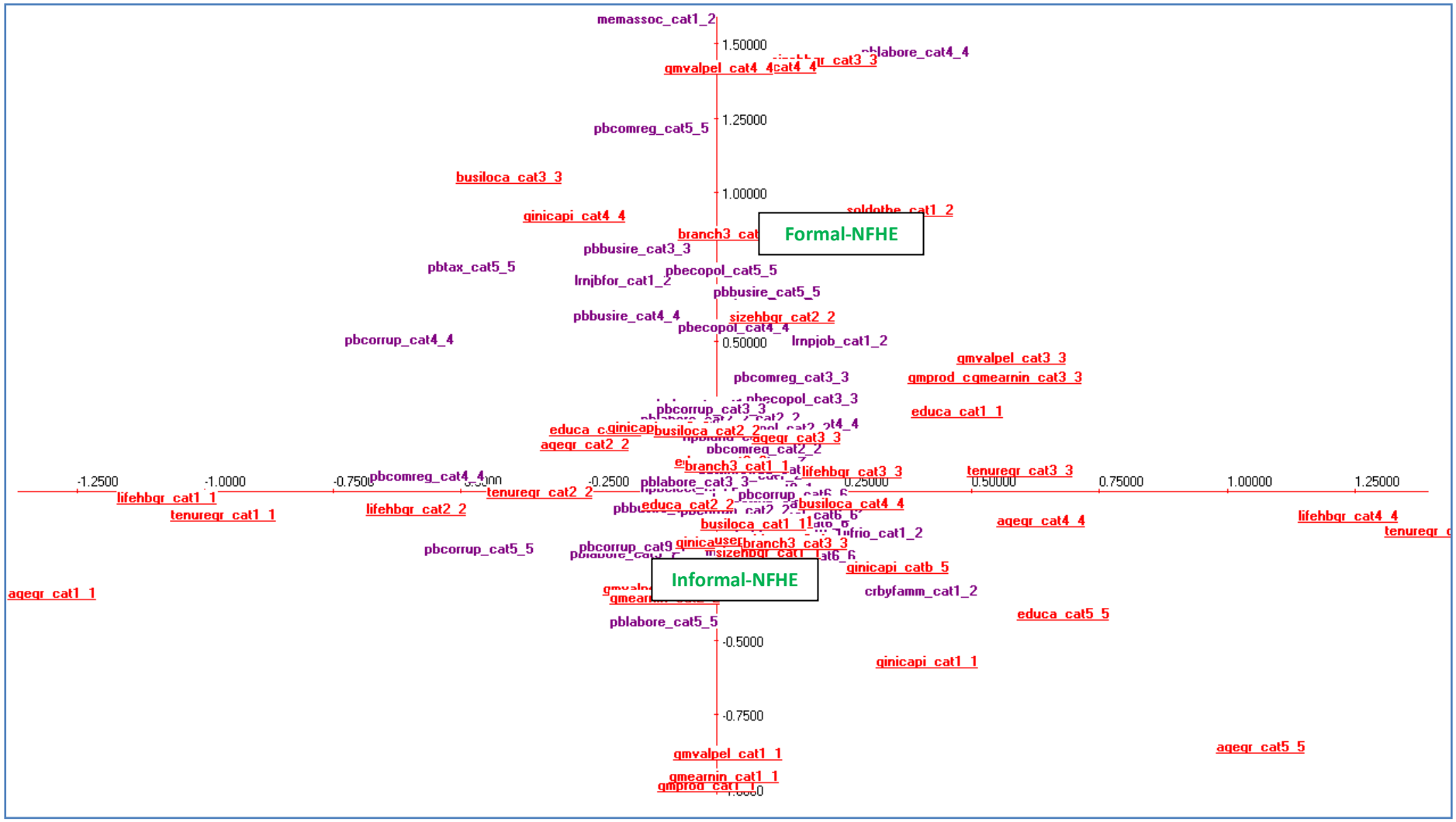
Aslo, the two variables tenure of the head of RNFHEs and firm life do not contribute entirely to the third facturial axis. They also participate in characterizing the fifth axis (Tables 5.11), with the opposition of modalities of these two variables in the same way as they are represented for the first factorial axis.

**Table 5.11: Absolute contributions and squared cosines of variables and associated modalities of responses that contribute to the fifth factorial axis**

| Variables/modalities | Sign of coordinate | Absolute contributions | Squared cosines |
|---|---|---|---|
| Tenure of the head of NFHE (<3 years) | - | 15.8 | 0.33 |
| Tenure of the head of NFHE (3-5 years) | - | 8.6 | 0.19 |
| Tenure of the head of NFHE (6-10 years) | + | 7.4 | 0.16 |
| Life of NFHE (<3 years) | - | 7.4 | 0.23 |
| Life of NFHE (6-10 years) | + | 19.0 | 0.56 |

*Source: VHLSS 2004, authors' calculation*

**Figure 5.12: plane of factors 1 and 3**



### 5.3. Typology of RNFHEs by cluster analysis

Based on the dimensional space generated by MCA, cluster analysis is then applied aiming at obtaining a typology of RNFHEs. Clusters of RNFHEs differ from each other along the factorial dimensions and also with respect to other characteristics of firms and owners.

Clustering based on K-mean like technique, followed by a hierarchical classification of the clusters provides 5 clusters of RNFHEs. Each cluster consists of a number of RNFHEs which are classified as being similar. This is a test of inference proposed by Lebart, Morineau and Piron (2006), with the null hypothesis assuming that the average coordinate of a cluster on a factor is equal to zero. This means that the profile of the cluster on this factor does not differ from the profile of the whole sample. If a cluster has a test value greater than or equal to 1.65 (the threshold of 95%) in absolute value, the null hypothesis is rejected because the cluster has a profile significantly different from the average profile.

As shown in Table 5.13, while the values associated with each v-tests for all the clusters on factors axis 1, 3, 4 and 5 are above 1.65 or below -1.65, which means that each class is sufficiently differentiated from those in each cluster is fall within these thresholds. We observe, however, that the profiles of clusters 1 and 5 on factorial axis 2 are not significantly different from that of the whole sample. This reflects that "sector of activities and business location" can not be used to discriminate the these two clusters.

**Table 5.13: The position of each cluster in the dimesional space**

coordinates and test-values on axes 1 to 5

| classes | | | | coordinates | | | | | test-values | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| iden - name | effec. | abs.w | disto | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| cut a of the tree into 5 classes | | | | | | | | | | | | | |
| aa1a- class 1 / 5 | 399 | 399.00 | .45 | 1.17 | .02 | -.22 | .11 | -.18 | 26.12 | .50 | -4.94 | 2.42 | -3.99 |
| aa2a- class 2 / 5 | 144 | 144.00 | 1.33 | 1.56 | .21 | -1.38 | -.89 | .48 | 19.46 | 2.67 | -17.17 | -11.05 | 5.96 |
| aa3a- class 3 / 5 | 494 | 494.00 | .33 | .17 | .07 | .90 | 1.00 | .30 | 4.34 | 1.90 | 22.95 | 25.44 | 7.68 |
| aa4a- class 4 / 5 | 502 | 502.00 | .34 | -.51 | -.14 | .71 | -1.27 | .11 | -13.17 | -3.59 | 18.46 | -32.72 | 2.72 |
| aa5a- class 5 / 5 | 479 | 479.00 | .47 | -1.09 | -.01 | -1.08 | .48 | -.42 | -27.24 | -.35 | -26.94 | 11.95 | -10.40 |

*Source: VHLSS 2004, authors' calculation*

Tables (from 5.14 to 5.18) show the lists of variables and associated modalities that contribute to charaterizinng each cluster. The second column in each table (v.test[7] - a measure of association between variables) allows us to distinguish the categories which are positively correlated with

[7] The test of inference used here is based on the concept of test (Lebart, Morineau and Piron, 2006). Its null hypothesis is that the proportion of respondents in a class holding a modality (eg, a particular feature) is equal to the proportion of the total sample holding the same modality. At a confidence level of 95% when the value of a test method exceeds 1.65, the null hypothesis is rejected: the proportion of respondents holding a class this method is significantly higher than the proportion of the total sample holding this modality.

the cluster (well represented) to those which are negatively correlated with class i. e. under-represented within it. Cla/mod indicates what share (percentage) of all individuals with this category is found in this class (this cluster, this category). Mod/cla indicates what share (percentage) of all individuals of this cluster have category.

*Cluster 1: RNFHEs of employers, high initial capital economic outcomes with certain attachment to formal social networks.* It is shown in Table 5.14 that there are many variables and associated modalies contributing to elucidating the nature of this cluster. This cluster is different from other ones in that it includes a cohort of employers since the majority (84.3%) of RNFHEs with paid workers is classified into this cluster. Also, it has significantly higher proportion of 'big' size RNFHEs than that of the whole sample. Particularly, about 69% of those RNFHEs having 3-5 workers and 86% of those with 6 workers or more belong to this cluster. It is observed also in this cluster the presence of entrepreneurs who have either undergone formal training the related to the working field (8% of all the entrepreneurs in the cluster) or received on-job-training (17.7% of all the ones in the cluster). Generally, the educational attainment of entrepreneurs in this cluster is rather high with about 45% of them have got upper secondary education, compared with 36% in the whole sample.

**Table 5.14: Characteristics of cluster 1**

```
-------------------------------------------------------------------------
                    characterizing with categories
  the classes or categories of:  cut a of the tree into  5 classes

-------------------------------------------------------------------------
 characteristic categories   v.test  --- percentage--- weight proba
                               cla/mod mod/cla global
-------------------------------------------------------------------------
        class   1 /  5                 19.08    385       aa1a

qmearnin_cat4_4        38.73  94.18   92.47   18.73    378 .0000 cat4
qmprod_cat4_4          31.60  79.50   82.60   19.82    400 .0000 cat4
qmvalpel_cat4_4        30.27  83.09   75.32   17.29    349 .0000 cat4
usepaidw_cat1_2        16.68  82.84   28.83    6.64    134 .0000 cat1
informal_cat0_1         9.42  40.82   31.17   14.57    294 .0000 cat0
sizehbgr_cat3_3         8.31  68.97   10.39    2.87     58 .0000 cat3
sizehbgr_cat2_2         7.69  38.28   25.45   12.69    256 .0000 cat2
sizehbgr_cat4_4         7.31  82.76    6.23    1.44     29 .0000 cat4
soldothe_cat1_2         6.71  42.47   16.10    7.23    146 .0000 cat1
lrnpjob_cat1_2          5.70  35.98   17.66    9.37    189 .0000 cat1
lrnjbfor_cat1_2         3.94  37.80    8.05    4.06     82 .0000 cat1
educa_cat3_3            3.67  23.43   44.68   36.37    734 .0001 cat3
memassoc_cat1_2         3.46  85.71    1.56     .35      7 .0003 cat1
pbbusire_cat3_3         3.42  48.28    3.64    1.44     29 .0003 cat3
pbtax_cat3_3            3.20  39.58    4.94    2.38     48 .0007 cat3
infotrad_cat1_2         3.09  21.16   70.13   63.23   1276 .0010 cat1
hpbtrans_cat1_2         3.09  22.85   42.08   35.13    709 .0010 cat1
agegr_cat3_3            2.92  22.84   39.22   32.76    661 .0018 cat3
hpbfinac_cat1_2         2.74  25.26   18.96   14.32    289 .0031 cat1
branch3_cat3_3          2.73  23.43   30.13   24.53    495 .0031 cat3
agegr_cat2_2            2.53  22.64   34.81   29.34    592 .0057 cat2
hpbpost_cat1_2          2.49  26.32   12.99    9.42    190 .0063 cat1
crbyfamm_cat0_1         2.41  19.80   92.99   89.59   1808 .0079 cat0
pbecopol_cat3_3         2.39  31.82    5.45    3.27     66 .0083 cat3
busiloca_cat3_3         2.37  54.55    1.56     .55     11 .0090 cat3
-------------------------------------------------------------------------
```

This cluster can also be charaterized by the highest concentration of formal RNFHEs. Within this cluster, there is a sub-group of 122 RNFHEs that are formally registered with business license. They account for 31.5% of total NFHEs classified into this cluster, and 41.5% of all the formal RNFHEs in the sample. These are also the most capital-intensive (with the mean of capital intensity of near 11 million VND) and are the most productive and profitables RNFHEs as well. Three quarter of the RNFHEs having the value added per capita and located in the highest quintiles are members of this cluster. Even informal RNFHEs in this cluster have much higer productivity and economic outcomes than their formal counterparts in other cluster. Products made by these RNFHEs serve not only local markets but also those of other provinces. It is reported that about 16% of RNFHEs in this cluster (vs. only 7.2% in the whole sample) sold their products to customers in other provinces. It is important, however, to note that though this is the cluster with the highest formality among the obtained clusters, the informal RNFHEs are prevailed with the share of about 67% of all RNFHEs in the cluster. These are the informal RNFHEs that have high performance and earnings.

It seems that with a larger scale of operation the RNFHEs are more likely to face some obstacles effecting the operation and development of their activity. Indeed, it is observed in this cluster the presence of RNFHE head who reported having difficulties related to communication, postal services (12.3%), transportation (42.1%) as well as financial access (18.9%). It is noted also that about 3 to 5% of the RNFHEs in this cluster declared having noteworthy hurdles concerning business registration, taxe and the inconsistancy of economic policies.

*Cluster 2: RNFHEs of own-account workers with moderate economic outcomes.* This cluster includes 395 RNFHEs (accounting for 19.6% of the total sample) with average firm life was about more than 8 years. Of all RNFHEs in this cluster, 96% had monthly earnings belong to the second highest quintile of the total sample. However, on average the RNFHEs in this cluster gererated monthly 851.733 VND (see Table A.2) of value added which is only one-forth of the average amount created by the RFHEs in the cluster 1.

**Table 5.15: Characteristics of cluster 2**

```
-----------------------------------------------------------------------
                  characterizing with categories
  the classes or categories of:  cut a of the tree into  5 classes

-----------------------------------------------------------------------
 characteristic categories   v.test  --- percentage--- weight proba
                                 cla/mod mod/cla global
-----------------------------------------------------------------------
       class   2 /  5               19.57    395       aa2a

qmearnin_cat3_3        38.15  86.14   95.95   21.80    440 .0000 cat3
qmvalpel_cat3_3        28.59  69.35   80.76   22.79    460 .0000 cat3
qmprod_cat3_3          21.34  57.85   65.32   22.10    446 .0000 cat3
usepaidw_cat0_1         3.88  20.44   97.47   93.36   1884 .0001 cat0
crbyfamm_cat0_1         3.21  20.52   93.92   89.59   1808 .0007 cat0
busiloca_cat2_2         2.90  25.20   23.80   18.48    373 .0019 cat2
sizehbgr_cat1_1         2.70  20.66   87.59   83.00   1675 .0034 cat1
-----------------------------------------------------------------------
```

Most of RNFHEs (87.6%) in this cluster is operated by one person, aged 39 years old on average, with the average years of schooling was more than 7 years. Another distinctive characteristic of this cluster is the higher proportion of RNFHEs operated in unfixed location than the average one reported for the whole sample. Although having a litle bit the same average level of capital-to-labour ratio in comparison with cluster 3, the RNFHEs in this cluster have relatively higher business outcomes and this can be explained by the difference between the two cluster in entrepreneurial characteristics and the frequence of activity. It is possible that the longer tenure and firm life as well as less casual activity of the RNFHEs in this cluster make them more productive. Overall, this is the cluster of RNFHEs that has the least distinctive characteristics from the mean values of the charateristics describing the total sample. In other words, this can be considered the medium cluster of all the obtained clusters.

*Cluster 3: Informal RNFHEs of young own account workers, low economic outcomes.* Of all the clusters this is the smallest one including 230 RNFHEs, accounting for 11.4% of the NFHEs in the sample. The most typical characteristics of this cluster is the higher rate of informality as well as a lower level of average working time in comparison with above clusters. About 91% of the RNFHEs in this cluster are informal. These are enterprises of own-account workers since nearly 100% among them have only one worker. This cluster holds about 60% of young entrepreneurs aged less than 24 years old in the whole sample. This also make the mean age of entrepreneurs 31 years old, significantly lower than that of whole sample. Also, they have lower seniority (95% less than 3 years) in the field of business that they were working and these are also short firm-life enterprises (74% less than 3 years).

**Table 5.16: Characteristics of cluster 3**

```
-----------------------------------------------------------------------
                  characterizing with categories
  the classes or categories of:  cut a of the tree into  5 classes

-----------------------------------------------------------------------
 characteristic categories   v.test  --- percentage--- weight proba
                                 cla/mod mod/cla global
-----------------------------------------------------------------------
       class   3 /  5                11.40    230       aa3a

tenuregr_cat1_1        25.41  46.50   95.22   23.34    471 .0000 cat1
lifehbgr_cat1_1        24.46  68.40   74.35   12.39    250 .0000 cat1
agegr_cat1_1           15.90  57.06   43.91    8.77    177 .0000 cat1
qmvalpel_cat2_2         4.48  16.76   40.87   27.80    561 .0000 cat2
usepaidw_cat0_1         4.28  12.10   99.13   93.36   1884 .0000 cat0
qmprod_cat2_2           3.98  16.19   39.13   27.55    556 .0000 cat2
qmprod_cat1_1           3.76  15.58   41.74   30.53    616 .0001 cat1
qmearnin_cat1_1         3.70  15.37   43.04   31.91    644 .0001 cat1
qmearnin_cat2_2         3.54  15.65   37.83   27.55    556 .0002 cat2
lrnpjob_cat0_1          3.46  12.14   96.52   90.63   1829 .0003 cat0
soldothe_cat0_1         3.38  12.02   97.83   92.77   1872 .0004 cat0
qmvalpel_cat1_1         3.34  14.97   42.17   32.11    648 .0004 cat1
informal_cat1_2         2.49  12.12   90.87   85.43   1724 .0064 cat1
-----------------------------------------------------------------------
```

The activities most strongly represented in this cluster include food and beverage (12.9%), textile, leather and wood processing (10.7%) and retail sale (44%). .

Regarding business outcomes, it is revealed that this cluster is rather heterogenous with the presence of RNFHEs having value added per labour and earnings ranged between the first and the second quintiles of associated distributions. On average an informal RNFHE in this cluster generated monthly 519,766 VND, 11% lower than their formal counterparts in the same cluster (see Table A.2).

*Cluster 4:. RNFHEs operated commerce activities with professional premises.* There are 461 RNFHEs in this cluster, about 22.8% of the sample. The most identifying characteristics of this cluster is the less precariousness in working condition since about more than one-third of RNFHEs in this cluster have a fixed business location which was separate from their home. This cluster can also be characterized by a high representation (56% vs. 45% in the total sample) of commerce RNFHEs of own account workers, particularly those who are retail sellers (51%). Resembling to cluster 3, almost all (99%) of these NFHEs do not have hired laborers and their average number of workers is 1.1 which was the lowest level among all cluster. The entrepreuneurs in this cluster were most likely those aged between 35 and 44 years old (38%) with tenure in the working activity ranging from 3 to 10 years (67%). The rate of informality among this cluster is not significantly different from the global rate which is reported of about 85% (see Table A3).

**Table 5.17: Characteristics of cluster 4**

```
-----------------------------------------------------------------------
                    characterizing with categories
  the classes or categories of:  cut a of the tree into  5 classes

-----------------------------------------------------------------------
 characteristic categories   v.test  --- percentage--- weight proba
                                 cla/mod mod/cla global
-----------------------------------------------------------------------
       class   4 /  5                 22.84    461       aa4a

qmearnin_cat2_2        36.77  78.78   95.01   27.55    556 .0000 cat2
qmvalpel_cat2_2        29.00  67.91   82.65   27.80    561 .0000 cat2
qmprod_cat2_2          20.66  55.58   67.03   27.55    556 .0000 cat2
sizehbgr_cat1_1         6.32  25.37   92.19   83.00   1675 .0000 cat1
usepaidw_cat0_1         5.90  24.15   98.70   93.36   1884 .0000 cat0
branch3_cat2_2          4.93  27.92   55.97   45.79    924 .0000 cat2
busiloca_cat4_4         4.66  30.26   35.57   26.86    542 .0000 cat4
tenuregr_cat3_3         2.67  27.04   32.32   27.30    551 .0038 cat3
tenuregr_cat2_2         2.62  26.69   35.14   30.08    607 .0044 cat2
agegr_cat3_3            2.52  26.32   37.74   32.76    661 .0058 cat3
lifehbgr_cat3_3         2.41  26.52   33.19   28.59    577 .0080 cat3
-----------------------------------------------------------------------
```

*Cluster 5: Newly startup Informal NFHEs working at home, low economic outcomes.* This cluster hold about more than 50% of all the old entrepreneurs aged 55 years old or more and 34% of those aged between 45 and 54 years old in the sample. This also make the mean age of entrepreneurs abou 41 years old, significantly higher than that of whole sample. Though on average the entrepreneurs in this cluster are old, they have lower seniority (less than 3 years) in

the field of business that they were working and these are also short firm-life enterprises (less than 3 years). Manufacturing are rather highly concentrate in this cluster, with the share of 42% of RNFHEs in the cluster and hold more one-third of all the manufacturing activities in the whole sample. With the high concentration of manufacturing the cluster is also characterized by a high proprotion (67%) of RNFHEs operating right at home. However, these are very small scale manufaturing enterprises since the cluster ranks lowest on the capital-intensiveness scale (see Table A1). In many respects this cluster of 479 RNFHEs resembles the cluster 4 described above. RNFHEs in this cluster have the lowest average productivity, earnings and other economic outcomes. An important feature that this cluster shares the common with the cluster 4 is the overwhelming of informal RNFHEs, accounting for about 95%

Regarding the perception of RNFHEs on business climate, contrary to what have been observed from the cluster 1, it is revealed that rather important proportion (about more than 60 %) of RNFHEs in this cluster confirmed that they did not face any institution-related difficulties concerning corruption and economic policies.

**Table 5.18: Characteristics of cluster 5**

characterizing with categories
the classes or categories of: cut a of the tree into 5 classes

| characteristic categories | v.test | --- percentage--- cla/mod | mod/cla | global | weight | proba | |
|---|---|---|---|---|---|---|---|
| class 5 / 5 | | 27.11 | 547 | aa5a | | | |
| qmearnin_cat1_1 | 40.04 | 82.76 | 97.44 | 31.91 | 644 | .0000 | cat1 |
| qmvalpel_cat1_1 | 36.36 | 78.40 | 92.87 | 32.11 | 648 | .0000 | cat1 |
| qmprod_cat1_1 | 29.47 | 71.59 | 80.62 | 30.53 | 616 | .0000 | cat1 |
| informal_cat1_2 | 7.73 | 30.05 | 94.70 | 85.43 | 1724 | .0000 | cat1 |
| agegr_cat5_5 | 7.65 | 51.79 | 18.46 | 9.66 | 195 | .0000 | cat5 |
| branch3_cat1_1 | 7.33 | 38.56 | 42.23 | 29.68 | 599 | .0000 | cat1 |
| busiloca_cat1_1 | 7.25 | 33.70 | 67.28 | 54.11 | 1092 | .0000 | cat1 |
| usepaidw_cat0_1 | 7.21 | 28.77 | 99.09 | 93.36 | 1884 | .0000 | cat0 |
| crbyfamm_cat1_2 | 5.75 | 44.76 | 17.18 | 10.41 | 210 | .0000 | cat1 |
| soldothe_cat0_1 | 4.13 | 28.21 | 96.53 | 92.77 | 1872 | .0000 | cat0 |
| hpbtrans_cat0_1 | 4.10 | 30.10 | 72.03 | 64.87 | 1309 | .0000 | cat0 |
| educa_cat5_5 | 4.01 | 35.60 | 24.86 | 18.93 | 382 | .0000 | cat5 |
| sizehbgr_cat1_1 | 3.91 | 28.84 | 88.30 | 83.00 | 1675 | .0000 | cat1 |
| loannetw_cat0_1 | 3.90 | 28.02 | 97.44 | 94.25 | 1902 | .0000 | cat0 |
| lrnjbfor_cat0_1 | 3.48 | 27.79 | 98.35 | 95.94 | 1936 | .0003 | cat0 |
| tenuregr_cat2_2 | 3.46 | 32.45 | 36.01 | 30.08 | 607 | .0003 | cat2 |
| agegr_cat4_4 | 3.36 | 34.10 | 24.50 | 19.47 | 393 | .0004 | cat4 |
| pbcorrup_cat1_1 | 2.51 | 29.29 | 61.79 | 57.19 | 1154 | .0061 | cat1 |
| lifehbgr_cat2_2 | 2.43 | 30.76 | 35.65 | 31.42 | 634 | .0076 | cat2 |

## 6. Conclusion

This study provides empirical evidence on important characteristics of the RNFHE sector in Vietnam which has not been sufficiently investigated in the literature. Our estimation shows a high rate of informality among RNFHEs in Vietnam. Regions in the north of Vietnam such as the Red River Delta and the Northeast tend to have higher rate of informal RNFHEs. On average, about nine in ten rural household businesses in these regions run an informal business activity.

Multiple correspondence analysis indicate significant loadings of informality, size, initial capital, and business outcome of RNFHEs in the first dimension. The results imply that RNFHEs in Vietnam are clearly distinct from one another in terms of informality. Another important and significant dimension seen as a source of heterogeneity among RNFHEs is made up of variables relating to firm age and actual experience of entrepreneurs. The cluster analysis divides the sample of RNFHEs into 5 groups which are distinguished by the underlying factors obtained from MCA. This analysis shows that the RNFHEs in Vietnam are found to be highly diverse in many aspects such as firm size, capital and human resources, operating conditions and business outcomes. In line with findings in some other empirical studies in developping countries (see for example Foti et al., 2007), the capital intensity of rural RNFHEs in Vietnam is also the major driving force in determining business outcomes.

As each obtained cluster is supposed to include homogeneous RNFHEs along the factorial axis, the results show the presence of both formal and informal RNFHEs in each cluster, implying that there is no evidence of duality among the RNFHE sector. A large number of informal RNFHEs is found in lower end of spectrums of business outcome indicators, but there appears also the presence of formal RNFHEs in this location. Also, we found in the cluster 1, where concentrated the big RNFHEs having high capital-intensity and business outcomes, the appearance of informal RNFHEs that generated high value added and income at much higher average levels than that of formal RNFHEs in other clusters. It follows from this that informal RNFHEs should not always be considered as a subsistence sector and that many formal RNFHEs are not so much different from their counterparts in the informal sector.

**References**


Benzécri, J. P. (1979). L'analyse des donnees. Tome 2 : l'analyse des correspondances, Dunod, Paris.

Cling, J.-P., Nguyen, T. T. Huyen, Nguyen, H.C., Phan T.N. Tram, Razafindrakoto, M. et Roubaud, F. (2010), "*The Informal Sector in Vietnam: A focus on Hanoi and Ho Chi Minh City*", Hanoi: Editions The Gioi. 190 p.

Cunningham, W. V. and Maloney, W. F. (1998), "Heterogeneity among Mexico's Micro-Enterprises: An Application of Factor and Cluster Analysis", Policy Research Working Paper No. 1999, Poverty Reduction and econimic Management Sector Unit, The World Bank.

Fafchamps M. and Minten B. (1998) "Return to social capital among traders", MSSD Discussion Paper No 23. , Market and structural Study Division, International Food Policy Research Institute.

Fetzer J. (1998) "Who is likely to be come self-employed in Vietnam?", Available at SSRN: http://ssrn.com/abstract=1161152

Foti, R., Mlambo, N., Muringai, V., Mavunganidze, Z. and Mapiye, C. (2007) "Heterogeneity among Zimbabwe's rural micro-enterprises and the determinants of their inclination to agro-dealership" , African Journal of business Management, pp. 1-7, February.

Gong, X., van Soest, A. and Villagomez, E. (2004). "Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico", Economic Development and Cultural Change, The University of Chicago, 2004.

Greenacre, J. (1984). Theory and applications of correspondence analysis. Academic Press.

Huang, Z. (1997) "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Set in Data Mining", In Research Issues on Data Mining and Knowledge Discovery (1997), pp. 1-8.

Jackle, A. E and Li, C. A. (2003) "Firms Dynamics and Institutional Participation: A case Study on Informality of Micro-enterprises in Peru"

Jovanovic, B. (1982) "Selection and the evolution of industry", Econometrica, Vol.50, No. 3, May.

Lebart, L., Morineau, A., and Piron, M., (2006). Statistique Exploratoire Multidimensionnelle, Dunod, Paris.

Leveson, A. R. and Maloney, W. F. (1988). "The Informal Sector, Firm Dynamics, and Institutional Participation", Policy Research Working Paper, The World Bank.

Livingstone, I. (1991). "A reassessement of Kenya's Rural and Urban Informal Sector", World Development, Vol. 19, No. 6. pp. 651-670, 1991.

Lucas, R. E., Jr. (1978) "On the size distribution of Business Firms" Ell Journal of Economics, 9:2, 08-23

Maloney, W. (1999). "Does Informality Imply Segmentation in Urban Labor Market? Evidence from Sectoral Transition in Mexico", World Bank Economic Review 13, No. 2.

Miles, N. and Norcliffe, G. (1984) "An Economic Typology of Small-Scale Rural Enterprises in Central Province, Kenya". *Transactions of the Institute of British Geographers*, New Series, Vol. 9, No. 2 (1984), pp. 211-229. http://www.jstor.org/stable/622169

Meyer, L. R. (1992), "Supporting Rural Non-farm Enterprises: What can be leaned from donor programs", Journal of Philiippine Development, No 35, Vol. XIX, Second Semester. P. 114-136.

Oostendorp, R. H., Tran, Q. T. and Nguyen, T. T. (2006) "The Changing Role of Non-farm Household Enterprises in Vietnam"

Rufer, T. and Knight, J. (2007) "Informal Sector Labour Market in Developping Countries", Oxford Policy Management, February 2007.

Vijverberg, W. (1988) "Non-agricultural Enterprises in Cote d'Ivoire: A descriptive Analysis", Living Standards Measurement Study, Working Paper No. 46, April.

Vijverberg, W. and Houghton, J. (2004) "Household Enterprise in Vietnam: Survival, Growth and Living Standards", in Economic Growth, Poverty, and Household Welfare in Vietnam (P. Glewwe, N. Agrawal, and D. Dollar, eds). Washington D.C.: World Bank.

Vu T. A. (2006), "Off-farm Business as a Poverty Reduction Actor in Vietnam Rural Areas", ASEDP No.73, Institute of Developing Economies, Japan External Trade Organization.

Zuwarimwe, J. and Kirsten, J. (2007) "Social network and social learning among small-scale rural nonfarm enterprises: a case study in chimanimani district in Zimbabwe", Labour and Management in Development Journal, Vol.8.

## Appendix

### Table A.1: Statistical Description of clusters

| Variables | Clust__1 | Clust__2 | Clust__3 | Clust__4 | Clust__5 | Total |
|---|---|---|---|---|---|---|
| Age of the head of RNFHE | 37.3 | 39.0 | 31.0 | 39.9 | 42.5 | 39.0 |
| Years of schooling of the head of RNFHE | 8.3 | 7.3 | 8.1 | 7.6 | 6.9 | 7.5 |
| Potential experience of the head of RNFHE | 24.0 | 26.7 | 17.9 | 27.3 | 30.6 | 26.4 |
| Tenure in the current job (in years) | 7.1 | 7.8 | 1.7 | 8.5 | 8.5 | 7.3 |
| RNFHE's life (in years) | 8.0 | 8.6 | 2.6 | 9.1 | 9.5 | 8.2 |
| Initial capital | 22468.6 | 5523.0 | 5384.5 | 3585.8 | 2017.9 | 7604.5 |
| Capital to labor ratio | 10624.7 | 4480.8 | 4449.7 | 2777.6 | 1434.8 | 4427.9 |
| Working days over last 12 months | 239.7 | 263.6 | 186.5 | 239.0 | 197.7 | 226.7 |
| Size of RNFHE | 1.7 | 1.1 | 1.2 | 1.1 | 1.1 | 1.2 |
| Number of RNFHEs with the same activity | 31.3 | 27.7 | 21.7 | 38.6 | 33.2 | 31.5 |
| Monthly production (1000 VND) | 5077.6 | 1232.3 | 725.9 | 732.8 | 375.6 | 1563.3 |
| Monthly intermediate costs (1000 VND) | 1835.8 | 382.4 | 204.1 | 243.1 | 167.5 | 550.2 |
| Monthly value added per labor (1000 VND) | 3241.8 | 849.9 | 521.8 | 489.7 | 208.0 | 1013.0 |
| Learning job through tutorial | 35.4 | 28.53 | 32.44 | 27.06 | 32.07 | 30.92 |
| Learning job through formal training course | 8.79 | 2.83 | 4.44 | 3.9 | 1.62 | 4.06 |
| Learning job by doing a paid job relating to current work | 18.35 | 9.77 | 3.11 | 7.36 | 7.03 | 9.37 |
| Particiapate in other types of business owned by household | 6.2 | 7.71 | 4.89 | 4.55 | 6.13 | 5.95 |
| RNFHE was created by another household member | 6.98 | 5.91 | 5.78 | 10.17 | 18.02 | 10.41 |
| RNFHE was created by friend or other aquaintence | 6.72 | 8.23 | 5.33 | 8.66 | 7.21 | 7.43 |
| Have relationship with other RNFHEs doing the same activity owned by relatives | 24.29 | 23.65 | 27.56 | 25.54 | 27.21 | 25.62 |
| Have relationship with other RNFHEs doing the same activity owned by friend | 45.99 | 45.76 | 38.22 | 40.69 | 38.74 | 41.87 |

**Table A2: Clusters of RNFHEs by economic activity**

| Activity | Clust_1 | Clust_2 | Clust_3 | Clust_4 | Clust_5 | Total |
|---|---|---|---|---|---|---|
| Food and beverage | 8.79 | 11.05 | 12.89 | 8.87 | 17.12 | 11.99 |
| Textile, leather, wood | 10.34 | 9.51 | 10.67 | 9.96 | 22.52 | 13.48 |
| Other manufacture | 9.04 | 3.08 | 3.56 | 1.3 | 2.52 | 3.72 |
| Construction | 2.07 | 0.26 | 0 | 0 | 0.18 | 0.5 |
| Whole sale | 9.82 | 2.57 | 2.67 | 3.68 | 2.16 | 4.11 |
| Retail sale | 29.97 | 44.73 | 44 | 51.08 | 38.92 | 41.67 |
| Hotel and restaurant | 11.63 | 10.28 | 8.89 | 9.09 | 4.68 | 8.57 |
| Transportation and whare house | 13.95 | 13.88 | 9.33 | 8.87 | 5.77 | 10.01 |
| Other services | 4.39 | 4.63 | 8 | 7.14 | 6.13 | 5.95 |
| Total | 100 | 100 | 100 | 100 | 100 | 100 |

**Table A3 : Heiterogeneity among formal/informal RNFHEs in terms of value added and income of RNFHE**

```
table     1     as row      : informal_V13                                          total weight:  2018.
              as column   : New_Partition
------------------------------------------------------------------------------------------------------------
      weight        ! Clust__1     ! Clust__2     ! Clust__3     ! Clust__4     ! Clust__5     !   together
% column            !              !              !              !              !              !
            %  row  !              !              !              !              !              !
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
                    !      120     !       66     !       21     !       58     !       29     !      294
informal_cat0_1     !  31.17       !  16.71       !   9.13       !  12.58       !   5.30       !  14.57
                    !       40.82  !       22.45  !        7.14  !       19.73  !        9.86  !      100.00
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
                    !      265     !      329     !      209     !      403     !      518     !     1724
informal_cat1_2     !  68.83       !  83.29       !  90.87       !  87.42       !  94.70       !  85.43
                    !       15.37  !       19.08  !       12.12  !       23.38  !       30.05  !      100.00
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
                    !      385     !      395     !      230     !      461     !      547     !     2018
      together      ! 100.00       ! 100.00       ! 100.00       ! 100.00       ! 100.00       ! 100.00
                    !       19.08  !       19.57  !       11.40  !       22.84  !       27.11  !      100.00
------------------------------------------------------------------------------------------------------------
 chi2 = 131.36 /  4 degrees of freedom
 0 theoretical frequencies < 5
proba ( chi2 > 131.36 ) =  .000 / test v =  10.79
------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------
table     2     as row      : informal_V13                                          total weight:  2018.
              as column   : New_Partition
                means  of : mvaladd_V43
------------------------------------------------------------------------------------------------------------
       weight       ! Clust__1     ! Clust__2     ! Clust__3     ! Clust__4     ! Clust__5     !   together
       mean         !              !              !              !              !              !
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
informal_cat0_1     !      120     !       66     !       21     !       58     !       29     !      294
                    !    4242.881  !     897.933  !     595.830  !     521.941  !     242.599  !    2102.822
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
informal_cat1_2     !      265     !      329     !      209     !      403     !      518     !     1724
                    !    2781.999  !     842.465  !     517.768  !     486.062  !     207.647  !     827.180
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
  together          !      385     !      395     !      230     !      461     !      547     !     2018
                    !    3237.339  !     851.733  !     524.895  !     490.576  !     209.500  !    1013.027
------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------
table     2     as row      : informal_V13                                          total weight:  2018.
              as column   : New_Partition
                means  of : mearning_V45
------------------------------------------------------------------------------------------------------------
       weight       ! Clust__1     ! Clust__2     ! Clust__3     ! Clust__4     ! Clust__5     !   together
       mean         !              !              !              !              !              !
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
informal_cat0_1     !      120     !       66     !       21     !       58     !       29     !      294
                    !    3496.456  !     853.038  !     583.619  !     486.375  !     232.661  !    1779.211
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
informal_cat1_2     !      265     !      329     !      209     !      403     !      518     !     1724
                    !    2493.676  !     835.154  !     519.766  !     483.564  !     208.730  !     781.450
--------------------+--------------+--------------+--------------+--------------+--------------+-----------
  together          !      385     !      395     !      230     !      461     !      547     !     2018
                    !    2806.230  !     838.142  !     525.596  !     483.917  !     209.999  !     926.812
------------------------------------------------------------------------------------------------------------
```