VNU Journal of Science, Earth Sciences 25 (2009) 65-75

Application of the Principal Component Analysis to explore the relation between land use and solid waste generation in the Duy Tien district, Ha Nam Province, Vietnam

Pham Van Cu[1,*], Philippe Charrette[2], Dinh Thi Dieu[1],
Pham Ngoc Hai[1], Le Quang Toan[3],

[1]*International Centre for Advanced Research on Global Change, VNU Hanoi*
[2]*Unviversité du Québec à Montréal.*
[3]*Institute of Space Technology, Vietnam Academy of Science and Technology VAST*

Received 09 July 2009; received in revised form 22 July 2009

Abstract. The paper presents and discusses the methodology used and the results obtained by the application of the Principal Component Analysis (PCA) on a set of socio-economical and land use data collected in the Duy Tien district (Ha Nam province), Vietnam. Objective of this study is to use PCA as a data reduction method to verify if a relation could be established between the quantities of waste generated in a region and its land use and socio-economical characteristics. Data was collected by a team from the Center for Applied Research in remote sensing and GIS (CARGIS) at the University of Sciences in Ha Noi. This study is part of the research Project "*Study the land use changes and its influences to the waste in rural sector of Duy Tien District based on Remote Sensing and GIS utilization*". The project is funded by Vietnam National University for the period 2007-2009.

Due to the limited availability of statistic data only three types of economic activity relation are preliminarily chosen for PCA to reveal which activity is the predominant for each commune: Non-farming income/Agriculture Dimensions, Development of the Tertiary Sector/Agriculture and Non-farming Income /Built-up zones expansion. The quantity of waste is than compared with the activity identified as predominant. All these results are than imported to GIS environment to give the cartographic presentation and to serve the future analysis.

Keywords: Principal Component Analysis; Waste; Land use change; Economic activity; GIS.

1. Introduction

The paper presents and discusses the methodology used and the results obtained by the application of the Principal Component Analysis (PCA) on a set of socio-economical and land use data collected in the Duy Tien district (Ha Nam province), Vietnam. Ha Nam province is a rural area located about 60 kilometers south of Ha Noi, the national capital of Vietnam. The Duy Tien district has 19 rural communes and two towns. The rural communes all together counted 32 617 households spread in 139 villages back to year 2006. At the same time, the average population of a Duy Tien district's commune was 7 195 persons.

* Corresponding author. Tel.: 84-913300970.
E-mail: pvchanoi@vnn.vn

Data was collected by a team from the Center for Applied Research in remote sensing and GIS (CARGIS) at the University of Sciences in Ha Noi. This study is part of the research Project *"Study the land use changes and its influences to the waste in rural sector of Duy Tien District based on Remote Sensing and GIS utilization"*. The project is funded by Vietnam National University for the period 2007-2009.

The main objective of our study was to use PCA as a data reduction method with the intention to verify if a relation could be established between the quantities of waste generated in a region and its land use and socio-economical characteristics. SPSS was the statistical software used to perform the PCA.

A first collection of analysis results is presented, described and discussed here in depth for application purpose. The components extracted for the case study describe two dimensions of the present situation of Duy Tien district: *level of importance of non-farming income* and *agriculture*. Two other series of results are also briefly outlined and discussed. The first case exposes once more two dimensions: the development of the *tertiary sector* and *agriculture*. Finally, two dimensions were as well extracted for the last case study: the *level of importance of non-farming income* and the *built-up zone expansion*.

The presentation of the results is mainly based on cartography of the factor scores produced as a result of the application of the PCA on the dataset.

2. Methodology and data

As mentioned above, objective of our study is to verify if there exists a relation between the quantities of waste generated in a region and its land use and socio-economical characteristics. This study question is based on the fact that the increase of quantity of waste is consequence of demographic and economic growth (Christian Zurbrügg 2002; Aurobindo Ogra 2003; Đào Thắm 2007). In the context of Duy Tien where the economic development level of 19 communes and 2 towns is quite different, it is important to evaluate the importance of certain key factors in their economic activities and to verify the relation of these driving factors with the waste quantity. Those relations are non farming income/agriculture, tertiary sector/agriculture and non farming income/built-up zone expansion. The statistic data we use in this paper are provided by the Department of Natural Resources and Environment and the Department of Agriculture of Duy Tien district.

In this study PCA is the main tool to seek such a linear combination of variables in which the variance extracted from the variables is maximal. It then takes away this variance from the model and tries finding a second linear combination which could explain the maximum proportion of the remaining variance, and the process continues until all the variance is extracted (Agilent Technologies 2005; M. McAdams and A. Demirci 2006). This is called the principal axis method and results in orthogonal (thus uncorrelated) dimension representing these driving factors which we use to analyze and interprete the statistic data of Duy Tien. This approach is widely used in land use analysis (Jan Peter Lesschen, Peter H. Verburg et al. 2005).

Performing PCA with help by SPSS was attempted here with the aim of reducing the large number of original variables available (more than 150) to a smaller number of factors for modeling and interpretation purposes. In Duy Tien study there is rather small number of cases in the available dataset (N=21 corresponding to 19 communes and 2 township). Therefore, only a reduced number of variables could be used at a time to allow the PCA to produce significant results.

The "sampling adequacy" measured thought the Kaiser-Meyer-Olkin (KMO varies from 0 to 1) statistic was also taken into consideration in the variable choice. We used SPSS to calculate a global KMO along with an individual KMO for each variable included in the PCA. It is generally recognized in the literature that overall KMO should be 0.60 or higher to proceed with any factor analysis, including PCA (Vines 2000). The individual KMO have been used to determine what variables to exclude from the analysis by dropping the variable with the smallest KMO and re-running the PCA until a satisfactory global KMO is obtained (Marketing Dept. SPSS Inc. 2000).

Once these mathematical constraints were fulfilled and an acceptable solution was reached, decisions had to be made regarding the factors to retain in the analysis. The main criteria used have been the Kaiser's rule i.e. the components retained were the one having eigenvalues strictly greater than 1. Emphasis was also put on the comprehensibility of the factors. In other words, the components kept were to those whose dimension of meaning was readily comprehensible in the scope of the research.

Finally the factor scores in tabular format where exported from SPSS to EXCEL and saved as a DBF (dBase IV format) file. The resulting dataset was imported into ArcGIS. A join was created between the administrative divisions (communes) geographic layer and this tabular dataset in order to spatially represent the outcome of the PCA and detect potential spatial distribution patterns. The symbology used to "spatialize" the factor scores was based on graduated colors which symbolizes the lower (< 0) factor scores by cold colors while warm colors accounted for the highest scores. The natural breaks ("Jenks") method proposed by ArcMap was uses create the classes. This method selects the class breaks that best group similar values and maximize the differences between classes. Each component was singly mapped using distinct ArcMap projects.

In the next paragraphs we will present the results of analysis of the three case studies and to shorten the text we will skip intermediary steps of calculation of such indicators as KMO measure, Chi Square Test, p value.

3. Results and interpretation

3.1. Case study 1: Non-farming income/ Agriculture Dimensions

All the data used for this case study are collected for the year 2006. The variables used to perform the analysis of this case study are summarized in Table 1 below.

Table 1. Description of variables used for analyzing Non-farming income/Agriculture Dimensions

Variable	Label (english)	Description
Bep_rom_cui	wood_cooker	Number of wood or straw cookers found in the commune.
@06Ho_CNTTCN	Industry_hh	Number of household involved in the industrial or small industry sector.
@06DT_lua	rice_area	Land area dedicated to rice crop (paddy field) [ha].
D_chuyen_dung	public_servive_area	Land are used for public infrastructure (e.g. roads) [ha].
@06Ho_CNXD	IC_income	Number of household with major income from industry or construction.
Agriarea*	agri_area_pc	Percentage of total area dedicated to agriculture [ha].

*This field was calculated based on existing variables @06Dat_SD (total agriculture dedicated area) and Dien_tich (total area).

The outcomes of the PCA computed by SPSS are in Table 2. Correlation Matrix[(a] below:

Table 2. Correlation Matrix[(a)]

		wood_cooker	Industry_hh	rice_area	public_servive_area	IC_income	agri_area_pc
Correlation	wood_cooker	1.000	-.012	.846	.490	.381	.378
	Industry_hh	-.012	1.000	.165	.647	.887	-.496
	rice_area	.846	.165	1.000	.711	.496	.394
	public_servive_area	.490	.647	.711	1.000	.804	-.142
	IC_income	.381	.887	.496	.804	1.000	-.297
	agri_area_pc	.378	-.496	.394	-.142	-.297	1.000

[a] Determinant = 0.001

Table 3. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.227	53.785	53.785
2	2.032	33.868	87.654
3	.404	6.735	94.389
4	.240	3.996	98.385
5	.068	1.128	99.513
6	.029	.487	100.000

Extraction Method: Principal Component Analysis

The Table 3 presents the eigenvalues calculated by SPSS showing that the first and the second components have eigenvalues greater than 1, i.e., 3.227 and 2.032 relatively. The first component provides 53.8% of the variance of the dataset while the second component takes 33.9% of the variance. Hence, those two first components represent almost 88% of the total variance existing in the data.

As per Kaiser's rule only the two first components were extracted by SPSS for the dataset. A *varimax* rotation was performed in order to make factor loadings of each variable to be more clearly differentiated by factor. An oblique *oblimin* rotation was also performed afterwards in order to generate the factor correlation matrix which displays the the Pearson's r coefficients between both components. Because this method looks after a non-orthogonal (oblique) solution, the purpose of is this operation was to verify if a potentially significant correlation exists between the components. The rotated component matrix along with the component plot allows distinguishing the two components extracted by the PCA as shown in Table 4.

Table 4. Reproduced Component Matrixes

Component Matrix[(a)]

	Component	
	1	2
public_servive_area	.931	-.053
IC_income	.914	-.316
rice_area	.779	.580
Industry_hh	.710	-.637
agri_area_pc	-.094	.867
wood_cooker	.637	.659

Extraction Method: Principal Component Analysis.
[(a)] 2 components extracted.

Rotated Component Matrix[(a)]

	Component	
	1	2
Industry_hh	.951	-.071
IC_income	.917	.307
public_servive_area	.770	.527
rice_area	.262	.935
wood_cooker	.103	.911
agri_area_pc	-.604	.630

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
[(a)] Rotation converged in 3 iterations.

The first component includes the variables related to non agriculture-related revenues. This component carries the number of households whose principal income comes from other workmanships such as manual labors related to construction. The number of household being active in small industries is also part of that component. It is also concerned by the infrastructures since it includes the variable *public_servive_area.* The existence of public infrastructures like paved roads may stimulate the development of the industrial sector which in turn provides non-farming revenues. Thus, this first dimension can be seen as a relative measure of the relative importance of non-farming incomes in a given community.

The second component gathers the variables that reflect agriculture-related way of living such the use of wood of (rice) straw fueled devices for cooking purpose. People tend to use wood of straw cooker where these resources are abundant. A high percentage of area dedicated to agriculture as well as the importance of the net area used for rice paddles are also indicators of high levels of agricultural activities. The second component can hence be seen as an indicator of the relative importance that the farming sector occupies in the local economy.

Factor Scores Mapping

Also called *component scores* in PCA, factor scores are estimations of the actual values of individual cases (observations) for the components. They are computed by taking the case's standardized score on each variable, multiplied by corresponding factor loading of the variable for the given factor, and sum these products.

The individual factor scores have been computed for component 1 "Non-farming income" and component 2 "Agriculture" and mapped to show the spatial distribution of the scores. The mapping also displays the estimated quantity of waste (as kilograms per person per month) generated in each commune. This aims to help visually perceiving the relationship (if any) between each extracted component and the waste production in the Duy Tien district. The maps are presented in the Appendix A.

"Non-farming income"

There are seven communes where the score for "Non-farming income" factor is greater than zero. Thus, in these communes (Hoàng Đông, Yên Bắc, Duy Minh, Mộc Nam, Chuyên Ngoại and Châu Giang) the non-farm income was more important than in the "average" conditions of the Duy Tuy district in 2006. This doesn't mean that the households in these communes did not gain any income from agriculture. This result solely means that compared to the rest of the district the seven communes count more households which were provided with a non agriculture related income. It should be noticed that national roads pass over five communes for whom the factor scores for this component are greater than zero.

Two communes colored in yellow show a factor score very close to zero for the first component. This is Yên Nam (-0.01760) and Mộc Nam (0.06425). They represent the modal situation of the district as per non-farm income. The remaining communes are represented in cold colors. There is relatively less households in these communes earning non-farming revenue that in the rest of the district. One can observe that these low non-farming income communes are mostly located in the southern part of the district.

"Agriculture"

The mapping of results for this dimension shows that intensive agriculture tends to concentrate in the north-central part of the district. There are seven communes with factor scores superior to 0 for the second component (Tiên Hiệp, Yên Bắc, Trác Văn, Châu Giang, Tiên Ngoại, Tiên Nội and Yên Nam). These are the communes where the agricultural sector is the most vigorous in the district based on the rice paddles surfaces and the percentage of the land relatively dedicated to agriculture. On the opposite, communes with a less prominent agricultural sector (as compared to the district's average situation) are located at the periphery.

3.2. Case study 2: Development of the tertiary sector / agriculture dimensions

Table 5. Description of variables used for case study 2

Variable	Label (english)	Description
@06Ho_XD	Construction_hh	Number of household involved in the construction sector.
@06Ho_CNTTCN	Industry_hh	Number of household involved in the industrial or small industry sector.
@06Ho_TN	Trade_hh	Number of household involved in trading.
@06Ho_Van_tai	Transport_hh	Number of household involved in transportation.
@06Ho_Dvu	Service_hh	Number of household involved in the service sector
@06DNN_Bqho	agri_land_per_hh	Surface of agricultural land per household [m^2]
Agriarea*	agri_area_pc	Percentage of total area dedicated to agriculture [ha]

The correlation matrix computed by SPSS show the correlations between the variables used for this case is reproduced as shown on Table 6 below:

Table 6. Correlation mtrix

		Industry_hh	Construction_hh	Trade_hh	Transport_hh	Service_hh	agri_land_per_hh	agri_area_pc
Correlation	Industry_hh	1.000	.105	.374	.375	.254	-.472	-.450
	Construction_hh	.105	1.000	.317	.673	.608	-.247	.069
	Trade_hh	.374	.317	1.000	.680	.563	-.442	-.309
	Transport_hh	.375	.673	.680	1.000	.713	-.343	-.100
	Service_hh	.254	.608	.563	.713	1.000	-.386	.072
	agri_land_per_hh	-.472	-.247	-.442	-.343	-.386	1.000	.531
	agri_area_pc	-.450	.069	-.309	-.100	.072	.531	1.000

[a] Determinant = .033

One can already see from the matrix that the variables Industry_hh, Construction_hh, Transport_hh and Service_hh relate to each other and tend to "cluster" to structure a distinct component. The calculated overall Kaiser's measure of sampling adequacy (KMO) was 0.703 which is quite acceptable in the conditions of the study. The two first principal components represent more than 71% of the variance of the dataset (47.76% and 23% relatively) as shown on Table 7.

Table 7. Quantity of information represented by principle components

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.344	47.765	47.765
2	1.639	23.412	71.176
3	.597	8.532	79.708

We have then extracted two components using the *varimax* rotation method. The Table 8 reproduced below displays the "loadings" of the variables on each component.

Table 8. Loading of vatriables on each component.

	Component	
	1	2
Transport_hh	.882	-.245
Service_hh	.882	-.108
Construction_hh	.829	.059
Trade_hh	.627	-.493
agri_area_pc	.161	.868
agri_land_per_hh	-.295	.757
Industry_hh	.195	-.743

The variables related to economic activities such as transport, construction, trade and other services sectors are grouped in the first component. This dimension obviously represents the level of importance of the tertiary sector in the Duy Tien district's economy. Also known as the service industry or service sector which does not involve the extraction of resources nor their transformation but is based on the provision of services to businesses as well as final consumers[(1)]. The remaining variables are all strongly related to the second component which tends to aggregate the variables that directly relate to farming or show a strong inverse relationship with it. This is the case for the number of family involved in small industries: one can assume that there is a clear inverse relationship between this variable and the importance of farming activities in the local economy.

Factor scores mapping

As for the first case study presented, the individual factor scores have been extracted for both components (development of tertiary sector and agriculture) and mapped to show the spatial distribution of the scores. The maps are provided in the Appendix B and commented here in details. The mapping also displays the estimated quantity of waste generated in each commune.

"Agriculture"

The second component shows a similar tendency in the scores' distribution then the one observed for the first case study presented (Non-farming income vs. Agriculture). Some differences are noticeable in the intensity of agricultural activity. These differences are partly due to the fact that the *quantile* method proposed by ArcMap was used this time as the classification method to create the graduated color symbology instead of the *natural breaks*. However, important part of this dissimilarity can moreover be explained by inherent factors found in the data such as population density, and by the choice of variables. For instance, Tiên Ngoại commune is getting the highest score mainly because it has the lowest population density of the district and thus the highest surface agricultural surface per household. Other examples are Yên Bắc and Châu Giang which were showing previously the highest scores on the agriculture dimension. This case study also points out that, when considering as well the role that the small industries are playing in these communities, the agricultural sector appears less important than expected at first glance. These pieces of information were not taken into account in the first case study.

"Development of the tertiary sector"

On the other hand, the mapping of first component's scores spatially illustrates and compares the development of the tertiary sector in Duy Tien. The highest scores centralize around both small towns especially TT Hoà Mạc. It is reasonable to think that the proximity of an urban center is a factor in the growth of business activities. This consideration seems to have a greater impact on the development of the tertiary sector that the proximity of a major road.

[(1)] Source: *Insee (Institut national de la statistique et des études économiques)*, France.

3.3. Case study 3: Non-farming income / Built-up zones expansion dimensions

Six variables used to perform the principal component analysis on this case study are described on Table 9Table 9.

Table 9. Description of variables used for case study 3

Variable	Label (english)	Description
TL_tang	pop_growth	Population growth rate in %.
@06Ho_CNXD	IC_income	Number of household with major income from industry or construction.
resident_area00_06	resident_area00_06	Variation in residential surface from 2000 to 2006 [%].
D_chuyen_dung	public_servive_area	Land area used for public infrastructure (e.g. roads) [ha].
@06DNN_Bqho	agri_land_per_pe	Surface of agricultural land per resident [m^2].
@06Ng_NLNTS	AFA_pc_income	Proportion of household whose major income is from agriculture, forestry or aquiculture. [%]

The principal component analysis performed had extracted more than 72% of the variance of the dataset as reported by SPSS in the Table 10 presented here.

Table 10. Quantity of information extracted by each component.

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.751	45.845	45.845
2	1.585	26.410	72.256
3	.769	12.824	85.079
4	.557	9.284	94.363
5	.192	3.197	97.560
6	.146	2.440	100.000

Using yet again the *varimax* rotation method, two components were extracted by SPSS. The table below displays the "loadings" of their respectively related variables on each component.

Table 11. Loading of variable on components 1 and 2 in case study 3

	Component	
	1	2
IC_income	.923	-.044
AFA_pc_income	-.862	-.254
public_servive_area	.821	-.165
pop_growth	.075	.766
resident_area00_06	-.223	.725
agri_land_per_pe	-.567	-.695

It appears from the rotated component matrix that the first dimension extracted reflects the "non-farming income" concentration of the Duy Tien communes as in the first case study previously discussed in details in this document. The second component extracted for this case holds the variables related to demographic (population growth) and land use change (positive variation in residential area and negative variation in agricultural land area per persons). This dimension represents the extension of the built-up areas i.e. the land covered by buildings and other man-made structures and activities[2].

Factor scores mapping

The maps are provided in the Appendix C. It is interesting to have a deeper look at the mapping of the factor scores of the second components (built-up expansion). The most noticeable built-up expansion doesn't necessary happens only on the outskirts of the two towns as one could normally expect. Surprisingly, high scores are present in some of the most off-centered communes such as Tiên Phong and

[2] As per the land cover categories proposed by the GTOS programme (http://www.fao.org/gtos/).

Đọi Sơn. For Tiên Phong, this situation can be explained two causes combined. This community possesses the greatest population growth of the district. This increase in population puts pressure on the residential area which has increased by 60% between years 2000 and 2006 while the average variation for the district was only 44%. Furthermore, it is a fairly small commune; consequently it has initially a rather small surface of land used for agriculture. Similar considerations regarding the strong population growth (1.17%) and an aggressive increase in the residential area (70%) can be applied to the Đọi Sơn commune.

4. Conclusion

By excluding TT Đồng Văn and TT Hoà Mạc, the two small towns of districts, the median monthly quantity of waste generated in the rural communes is 9 kilograms per persons while the average quantity is 11 kg. Almost 60% of the communes (4/7) where the factor scores for the "Non-farming income" dimension are positive show a waste quantity greater than the average.

However, one commune, namely Yên Bắc, has a fairly high factor score (0.8530) for this dimension but generates a rather low quantity of waste. It is interesting to note that this commune also shows the second highest factor score for the other dimension (agriculture). One possible explanation is that the residents of this rural commune are also migrant workers between the crops. They can spend few months outside their village each year working in the construction sector in the surrounding towns. The additional revenues earned from this seasonal work would explain why this commune bears a high score on the "non-farm income" component. On the other hand, the fact that these inhabitants are temporary living away from their village could explain the above average waste quantity generated in the commune.

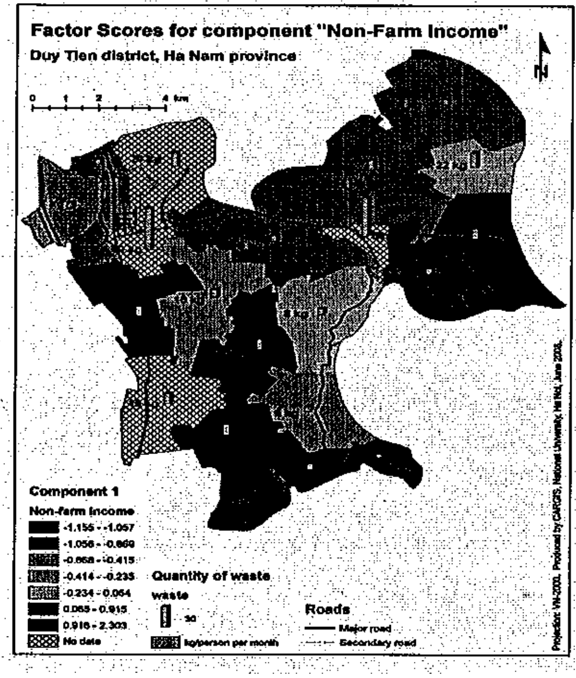
Mộc Nam commune shows an appreciatively average score for component "Non-farming income" and a very negative score on "agriculture" its waste generation is twice the average (22 kg). The craft sector (particularly handcrafted dye works) is well-developed in this off-centered commune. This typical activity could explain the more than expected waste generation of the commune.

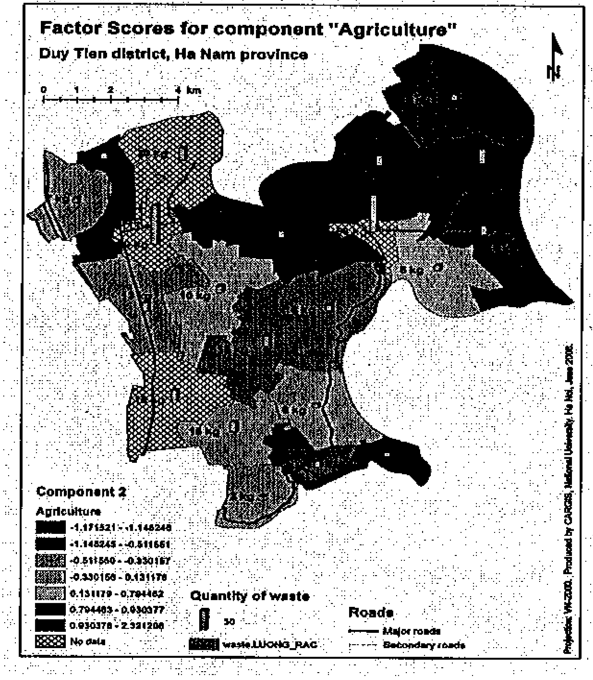
The proximity of a major road appears to have a positive impact on the level of non-farm income at least for communes located along by the main North-South and West-East axis roads. However, there is no obvious relation with the existence of a main road crossing a commune and the waste quantity generated locally.

In retrospect, it is apparent that based the information extracted from the Duy Tien data, the quantity of waste generated in a given locality is mostly determined by its function. The small towns of the district carry out commercial and industrial activities that are not present (or only at a much less intense level) in the rural communes. As a consequence of the role that the two towns play as trade centers, these localized activities generate more waste than any other activities originating from anywhere in the rest of the district.

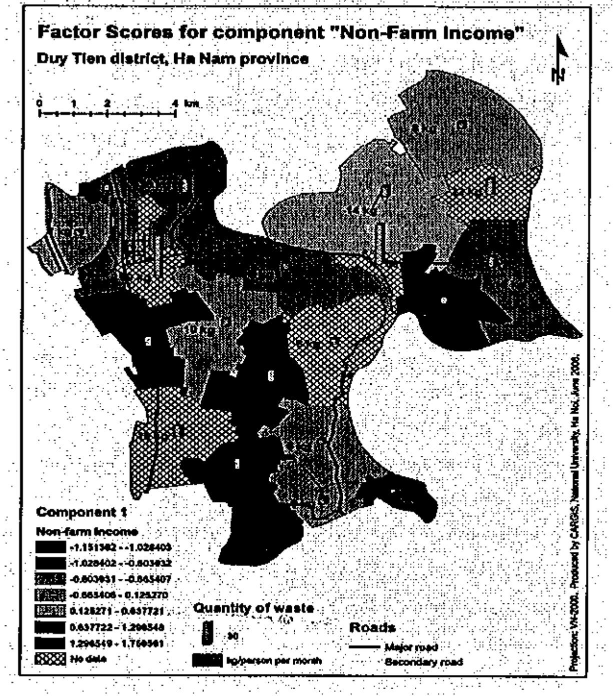
Minimally, there must be more cases than variables to perform a PCA. Many authors mention that factor analysis is inappropriate when sample size is below 50. Some arbitrary "rules of thumb" also exist and are widely used in practice to calculate the minimum number of cases required. For instance, according to Bryant and Yarnold (1995), the number of cases should be at least 5 times the number of variables entered in the analysis. It's essential to mention that the PCA reported here doesn't comply with this rule: only 21 cases were available while five to seven variables were included in the model.

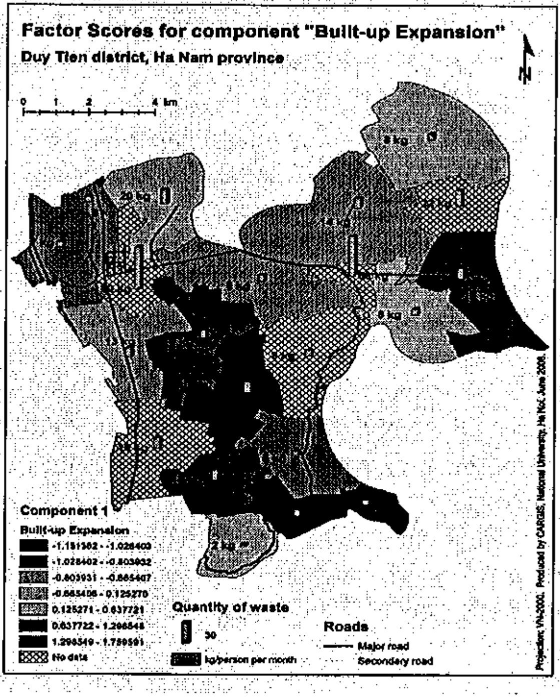
Appendix A - Maps of Factor Scores for *Non-farming income - Agriculture*



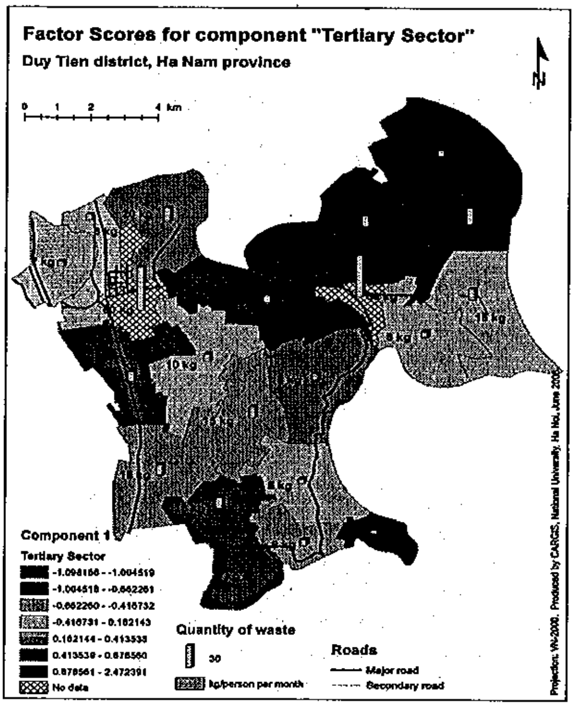


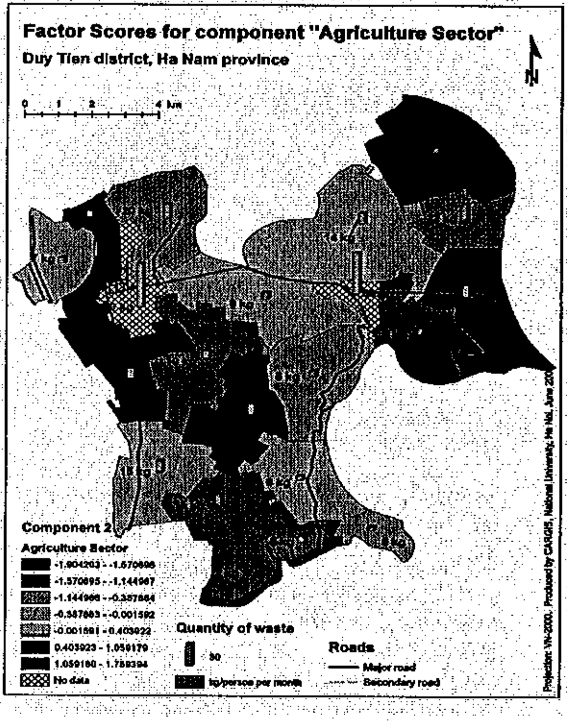
Appendix B - Maps of the Factor Scores for *Tertiary Sector - Agriculture*





Appendix C – Maps of the Factor Scores for Non-farming income - Built-up expansion





References

[1] I. Agilent Technologies, Principal Components Analysis, sig_support@agilent.com, 2005.

[2] Aurobindo Ogra. *Logistics Management and Spatial Planning for Solid Waste Management System using Geographic Information System Map Asia*, 2003.

[3] Christian Zurbrügg, *Urban Solid Waste Management in Low-Income Countries of Asia How to Cope with the Garbage Crisis*, Urban Solid Waste Management Review Session, Durban, South Africa, November, 2002.

[4] Đào Thắm, Về đâu rác thải sinh hoạt nông thôn? http://www.baohungyen.vn/content/viewer.asp?a=6848&z=63, 2007.

[5] Jan Peter Lesschen, Peter H. Verburg, et al, *Statistical methods for analysing the spatial dimension of changes in land use and farming systems*, The International Livestock Research Institute, Nairobi, Kenya & LUCC Focus 3 Office, Wageningen University, the Netherlands, 2005.

[6] M. McAdams, A. Demirci, *The use of principle component analysis in data reduction for GIS Analysis of water quality data,* Volume, DOI:, 2006.

[7] Marketing Dept. SPSS Inc., SPSS (Statistical Producers for Social Science), *SPSS software and manual*, Chicago, Illinois, USA, 2000.

[8] S. Vines, Simple principal components, *Applied Statistics* 49 (2000) 441-451.