Insight into apartment attributes and location with factors and principal components applying oblique rotation

Alain Bonnafous, Marko Kryvobokov, Pierre-Yves Péguy

Laboratory of Transport Economics (LET), Lyon, France

Abstract

Apartment characteristics including prices, internal attributes and location attributes consisting of travel times to urban centres and income variables are analysed with exploratory factor analysis. Principal axis factoring with oblique rotation is applied, which allows the extracted factors to be correlated. Four factors are extracted, of which two represent apartment attributes and other two – location attributes. The analysed area is the French adjacent cities of Lyon and Villeurbanne. Spatial distribution of the factors provides an insight into both apartment attributes and urban structure. In particular, factors show the concentration of big expensive apartments on the one hand and older apartments in bad condition on the other; they also demonstrate a contradiction with the existing city boundaries in the north and highlight the existence of a problematic low income area in the central part of Lyon. Principal component analysis is applied for a more comprehensive study of location attributes. The clusters of components obtained by K-means algorithm are seen as proxies for apartment submarkets, which are useful for a subsequent study.

Keywords: apartment attributes, location attributes, exploratory factor analysis, principal component analysis, oblique rotation.

1. Introduction

A complex social nature of real estate price is a well-known phenomenon. In the academic world the most popular way of its analysis is a hedonic regression modelling, where, in the cross-sectional version without focusing on time, the dependent variable is usually a price and the independent variables include real estate attributes and location attributes. The estimated parameters are interpreted as willingness to pay for different attributes (Rosen, 1974).

The other way of analysis does not imply focusing on price as dependent variable. The aim of such an analysis is a better understanding of data itself with insight into the hidden relationships between variables. The methods of this group include clustering, factor analysis, principal component analysis (PCA), artificial neuron networks and others. To higher or lesser degree the results of these methods are related to pattern recognition and can be applied for identification of neighbourhoods/submarkets/value zones and/or in hedonic regression.

A relatively often used technique is a combination of factor analysis or PCA and cluster analysis. The extracted factors or principal components are used as a data for clustering to determine submarkets and include them in hedonic price equation. For this purpose, Dale-Johnson (1982) applied Q-factor analysis, whereas Maclennan and Tu (1996), Bourassa *et al*. (1999), Bourassa *et al*. (2003) exploited PCA. For example, Bourassa *et al*. (2003) found that the best results were obtained when cluster analysis was based on the two most important components.

The other application of PCA in hedonic modelling of real estate prices was proposed by Des Rosiers *et al*. (2000). The mentioned study as well as Des Rosiers and Thériault (2008) use PCA in the Quebec Urban Community for data reduction. In particular, to avoid severe multicollinearity in hedonic price model induced by fifteen accessibility attributes of travel times and walking times to different objects, two principal components were obtained. Then these components were used in a regression model as substitutes for initial variables. The authors made a quite straightforward interpretation: the first component accounts for accessibility to regional services, while the second one refers to local accessibility. In the former study it was also obtained four principal components on census attributes. After mapping of the principal components, Des Rosiers *et al*. (2000) conclude that PCA provides useful insights into housing market dynamics: it clearly highlights the marked concentration of low income households dwelling as opposed to high-income households and also prove consistent with urban reality.

With the aim to identify latent construct underlying our variables, in this study we apply the methodology of exploratory factor analysis (EFA). According to Fabrigar *et al*. (1999) it is an appropriate form of analysis “if the goal is to arrive at a parsimonious representation among measured variables”. When the goal is data reduction, PCA can be applied (Bonnafous, 1973; Fabrigar *et al*., 1999). Though both methods represent the observed variables as linear combinations of factors or components and are closely related, they are not identical. PCA takes into account all variability in the variables, while factor analysis explains the variability, which exists due to common factors (“communality”, which in this case is less than unity).

The rotation method usually exploited in PCA applications in the real estate domain (e.g. by Bourassa and colleagues or Des Rosiers and colleagues) is a varimax rotation, which involves an orthogonal transformation of variables into a new set of mutually independent components. In the current study we apply an oblique rotation, which permits correlation among factors. As Fabrigar *et al*. (1999) noted, the methodological literature suggests little justification for using orthogonal rotation; it can be reasonable only if the oblique solution indicates that the factors are uncorrelated.

The paper is organised as follows. The subsequent section describes a non-routine process of data preparation for factor analysis. The third section deals with the EFA itself and includes the interpretation and geographical demonstration of factors. The fourth section is about the application of PCA to location attributes. In the penultimate section, the clusters of principal components are created, while the final section concludes.

2. Data preparation

Geographically the area of study includes the cities of Lyon and Villeurbanne. These adjacent cities with overall population of over 600 thousand inhabitants have a common planning structure and transportation network and make up the core of the Lyon Urban Area, which is the second largest agglomeration by population in France.

The data on sale prices and apartment attributes were provided by *Perval*, which collects information about real estate transactions in France. Data on approximately 10,000 apartment sales selected randomly from all sales in the central part of the Lyon Urban Area in the period of 1997-2008 were obtained. With very few exceptions, the apartments are located in the urbanised area and mainly concentrated in Lyon and Villeurbanne.

We deleted observations with missing data and with prices lower than 20,000 Euros and higher than 500,000 Euros and with area of less than 18 square metres and more than 200 square metres. We also deleted the observations, for which the standardised residuals of the linear OLS hedonic price model are higher than three (see the details in Kryvobokov, 2009). The 4,251 remained observations are used in the analysis. Exclusion of more observations with missing data could significantly lower sample size and the statistical power of results, while attributing mean scores for missing values reduces variation among observations and increases the potential for clumping and truncation (Vias and Kumaranayake, 2006). In our study, 26% of observations have no data about the number of parking places and 60% have no data about the quality of view. We choose to exclude these variables, because otherwise we would be enforced either to arbitrarily use mean scores or to considerably decrease our simple size.

Location of apartments is demonstrated in Figure 1, where the boundaries of IRISes are shown with thin lines and the boundary of Lyon and Villeurbanne is shown with thick line. IRIS (les îlots regroupés pour l'information statistique) is a French statistical unit used also as a transport analysis zone. The definitions of variables and descriptive statistics are presented in Tables 1-2, of which the former includes apartment attributes and the latter describes location attributes.

As factor analysis is designed for continuous data, we treat our count variables (e.g. number of rooms) and categorical variables (e.g. construction period) not as dummies, but as continuous variables (see Kolenikov and Angeles, 2004). Thus, there are seven construction periods: before 1850; 1850-1913; 1914-1947; 1948-1969; 1970-1980; 1981-1991; and 1992 and later, which are treated as continuous variables equal to 1 to 7 respectively, though we admit that this representation is rather artificial. The same situation is with the attribute of transaction year represented as the interval of integers of 1-12 corresponding to 1997-2008 and with the variable for state of apartment represented as 1, 2, and 3, which correspond to “renovation is needed”, “preventive maintenance is needed”, and “good state” respectively.

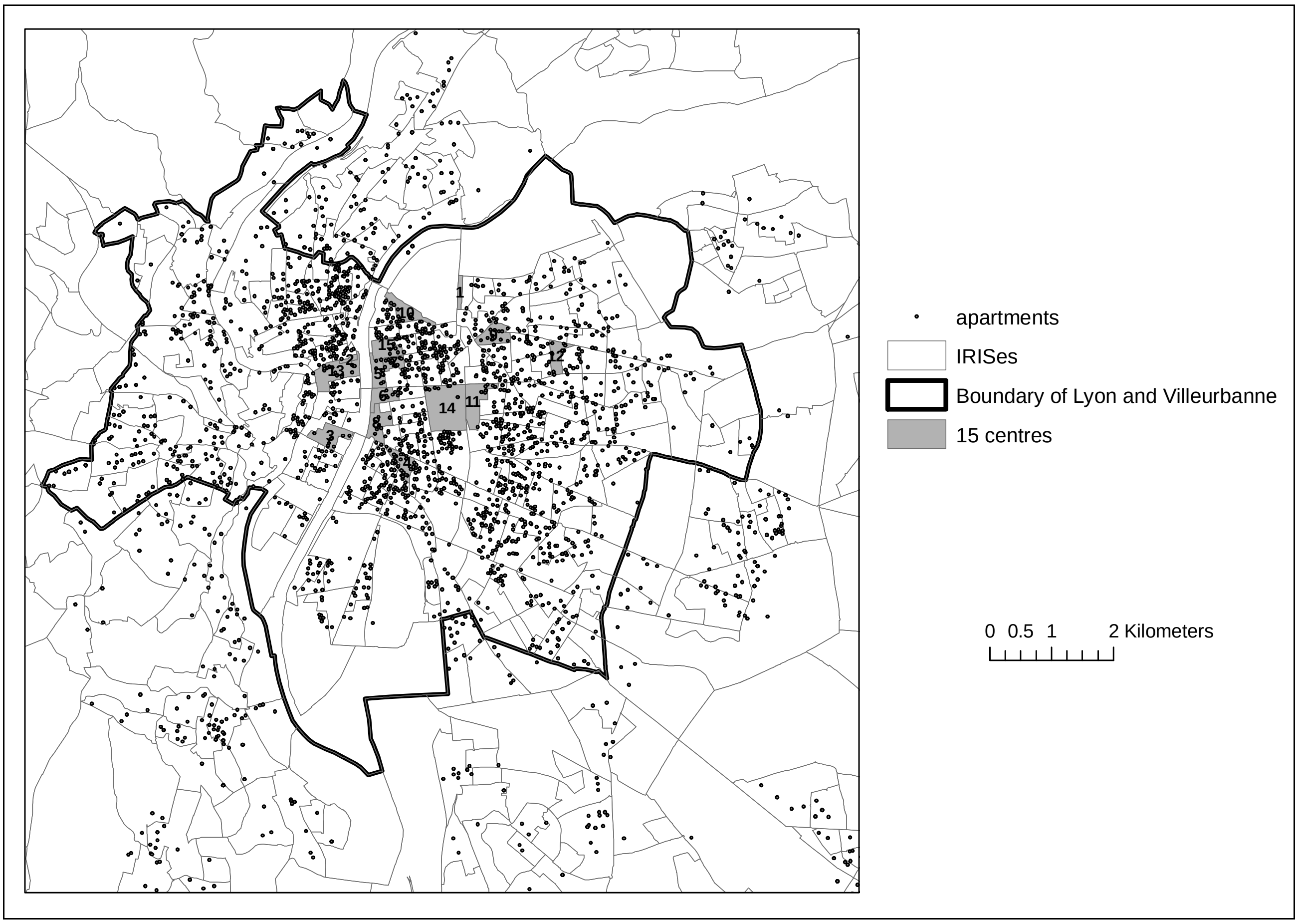

Figure 1. Location of apartments

Location variables in Table 2 include percentages of households in three income groups and travel times to urban centres in minutes. Both groups of location attributes are calculated per IRIS. The middle income group includes households in the middle 60% of the income range, and the lowest and highest 20% margins compose the other two groups. Travel times for the a.m. peak period by public transport for this study were obtained from the MOSART transportation model for the Lyon Urban Area. We take into consideration fifteen service employment centres, which were identified with residual analysis in Kryvobokov (2009). Location of service employment centres is presented in Figure 1.

Normality is checked with skew and kurtosis taking into account the thresholds of 2 and 7 respectively (West *et al.*, 1995). The highest skew (for *Condition*) is only a bit higher than 2, whereas kurtosis for all variables is lower than 4. Many available apartment attributes are not included in the analysis because of their severe non-normality. It refers to the number of bathrooms, the area of garden and others.

Table 1. Definition of apartment variables and descriptive statistics

Variable	Description	Mean	Minimum	Maximum	Std. deviation	Skew	Kurtosis
Price	Transaction price, Euros	122,235.90	20,276.00	500,000.00	69,979.67	1.45	2.93
Year_Sale	Count for year of transaction	6.87	1	12	2.87	-0.10	-0.88
Area	Apartment area, square metres	68.63	18	196	25.98	0.78	1.51
Rooms	Number of rooms	3.05	1	8	1.19	0.26	-0.18
Floor	Floor	2.84	0	18	2.25	1.35	3.85
Const_Period	Construction period	5.12	1	7	1.75	-0.50	-0.73
Condition	State of apartment	2.79	1	3	0.47	-2.14	3.87
Cellars	Number of cellars	0.69	0	2	0.50	-0.43	-0.88

Table 2. Definition of location variables and descriptive statistics

Variable	Description	Mean	Minimum	Maximum	Std. deviation	Skew	Kurtosis
%LowIncome	Percentage of low income households	29.42	10.24	52.12	5.78	-0.10	-0.05
%MidIncome	Percentage of middle income households	58.00	42.70	66.20	3.30	-0.15	0.09
%HighIncome	Percentage of high income households	12.58	4.34	28.77	2.92	0.51	0.68
TT_1	Travel time to Stalingrad	11.31	1.41	24.43	4.85	0.43	-0.25
TT_2	Travel time to Louis Pradel	11.18	2.22	29.36	5.35	0.62	0.01
TT_3	Travel time to Bellecour-Sala	10.99	0.45	31.28	4.96	0.89	0.79
TT_4	Travel time to Victor Bach	9.60	0.45	28.49	4.96	0.51	0.04
TT_5	Travel time to Molière	10.41	0.45	29.30	5.25	0.69	-0.02
TT_6	Travel time to Jussieu	10.44	0.45	30.36	5.18	0.72	0.01
TT_7	Travel time to Saxe-Bossuet	10.05	0.45	28.40	5.32	0.64	-0.19
TT_8	Travel time to Mutualité-Liberté	10.04	0.45	30.37	5.10	0.77	0.31
TT_9	Travel time to Charles Hernu	11.19	0.45	26.17	5.37	0.35	-0.66
TT_10	Travel time to Les Belges	11.00	0.45	27.48	5.34	0.49	-0.44
TT_11	Travel time to Villette Gare	10.68	0.45	29.25	5.35	0.37	-0.81
TT_12	Travel time to Gratte Ciel est	11.68	0.45	25.42	5.67	0.19	-0.79
TT_13	Travel time to Terreaux-Bat d’Argent	10.97	0.45	30.41	5.19	0.82	0.35
TT_14	Travel time to Part-Dieu	10.62	0.45	29.36	5.24	0.46	-0.71
TT_15	Travel time to Marechal Lyautey	10.39	0.45	28.46	5.29	0.65	-0.15

3. Factor analysis

Principal axes factoring is applied as the most widely used method in factor analysis (Warner, 2007). We use the standard method of non-orthogonal rotation – direct oblimin.

It was impossible to include in the analysis all the variables from Table 1 and Table 2. In particular, *Area* and *Rooms* could not be presented simultaneously, and the former variable was chosen. Surprisingly, *Year_Sale* and *Floor* demonstrated so low communality, that both attributes were excluded. The use of quarter of sale instead of year as well as calculation of trends for both year and quarter did not improve the situation. Of income groups, the two marginal ones were included.

Of fifteen variables of travel times to service employment centres, it was possible to include eight. Among the centres included are Bellecour-Sala usually referred to as the CBD and the two other commonly recognisable centres of Louis Pradel and Part-Dieu.

The communalities of the attributes presented in Table 3 are ranged from 0.99 to 0.14 with the mean of 0.75. It is problematic to find in the factor analysis literature the reported communalities referred for real estate. If to use for comparison the psychological studies reported in Fabrigar *et al*. (1999), who analysed data sets from Breckler (1984) and Crites *et al*. (1994), then our communalities are in general in line with them, though we should note that the number of observations in the mentioned sources is considerably less than in our case and our minimal values for *Condition* and *Cellars* are rather low. Nevertheless, we will keep both variables because they, especially the former one, represent important apartment attributes.

We select the number of factors using the criterion that the eigenvalues of the unreduced correlation matrix should be higher than one. There are four such factors; their eigenvalues are 7.32, 1.87, 1.60 and 1.22. The fifth factor has the eigenvalue of 0.84. The scree plot of eigenvalues (Figure 2) supports our choice: starting from the fifth factor, the slope becomes gentler. The factor correlation matrix (Table 4) shows that factor 1 and factor 4 are negatively correlated with the coefficient of higher than 0.50. Thus, the decision to apply a non-orthogonal rotation is right.

EFA analysis is a common factor model, where each measured variable is a linear function of one or more common factors (that influence more than one measured variables) and one unique factor (that influence only one measured variable). Factor loadings for structure matrix and pattern matrix are presented in Table 3. The first matrix represents the variance in a measured variable explained by a factor in both a unique and common contributions basis. The pattern matrix represents only unique contributions.

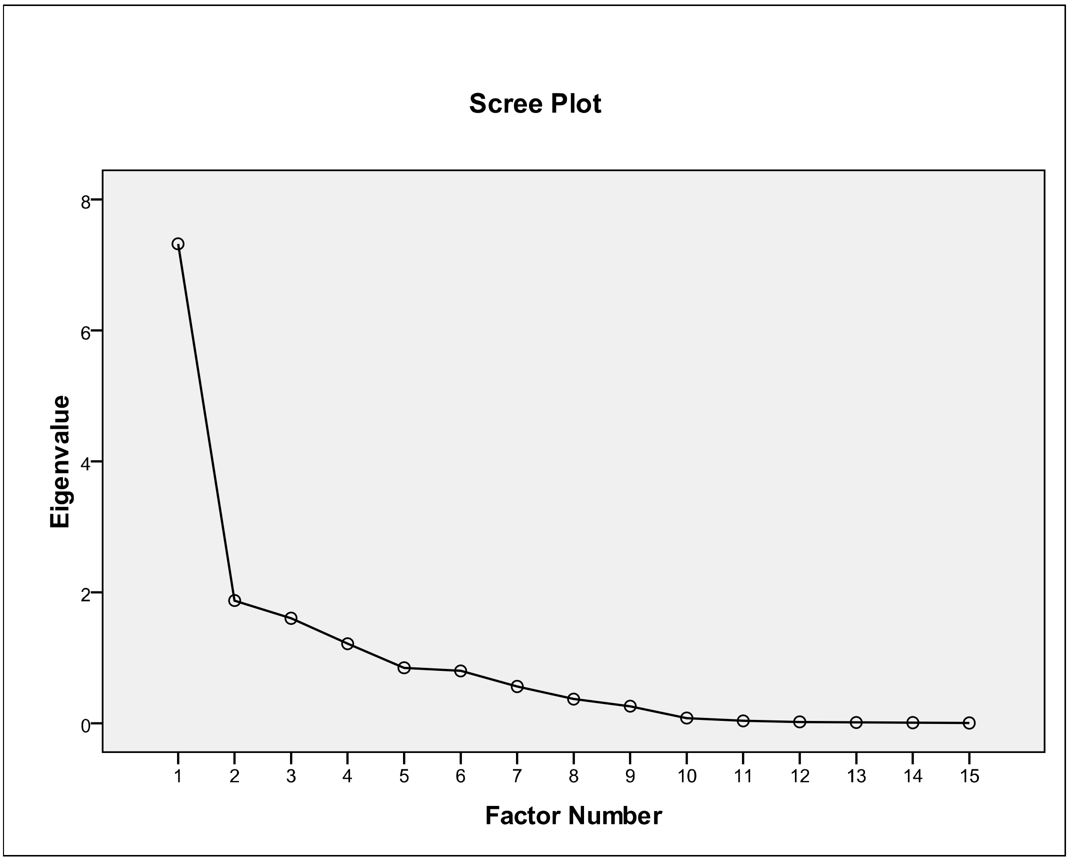

Figure 2. The scree plot of eigenvalues

Table 3. Communalities and factor loadings

Variable	Communality	Factors							
		Structure matrix				Pattern matrix			
		1	2	3	4	1	2	3	4
Price	0.56	-0.18	**0.86**	-0.08	-0.05	-0.12	**0.86**	-0.12	<0.01
Area	0.53	0.03	**0.82**	0.07	-0.13	0.10	**0.83**	0.04	0.02
Const_Period	0.34	0.08	0.04	**-0.78**	-0.13	0.01	0.06	**-0.77**	-0.08
Condition	0.14	0.02	0.08	**-0.40**	-0.04	<0.01	0.09	**-0.41**	-0.01
Cellars	0.18	0.04	0.18	**0.37**	-0.12	0.01	0.14	**0.36**	-0.11
%LowIncome	0.85	**-0.49**	-0.12	0.01	**0.93**	-0.01	<-0.01	-0.04	**0.93**
%HighIncome	0.86	**0.50**	0.10	-0.02	**-0.94**	0.03	-0.01	0.03	**-0.93**
TT_3	0.96	**0.68**	-0.07	-0.22	**-0.60**	**0.49**	-0.07	-0.18	**-0.34**
TT_10	0.98	**0.95**	-0.12	-0.15	**-0.48**	**0.95**	-0.05	-0.10	0.01
TT_6	0.99	**0.94**	-0.09	-0.17	**-0.63**	**0.82**	-0.05	-0.12	-0.20
TT_14	0.99	**0.95**	-0.02	0.04	**-0.54**	**0.92**	0.03	0.09	-0.07
TT_2	0.98	**0.87**	-0.14	-0.28	**-0.55**	**0.77**	-0.09	-0.23	-0.15
TT_9	0.98	**0.93**	-0.04	0.06	**-0.41**	**>0.99**	0.04	0.10	0.11
TT_11	0.99	**0.91**	-0.00	0.09	**-0.52**	**0.90**	0.05	0.14	-0.05
TT_1	0.96	**0.88**	-0.07	-0.00	**-0.37**	**0.95**	0.01	0.04	0.12

Table 4. Correlation between factors

Factor	1	2	3	4
1	1.00	-0.08	-0.05	-0.52
2	-	1.00	0.05	-0.12
3	-	-	1.00	0.05

We will focus on loading higher 0.30 and lower -0.30, which are in bold in Table 3. It is clearly seen that factor 1 and factor 4 are location factors, whereas factor 2 and factor 3 are the factors of apartment attributes. Significant difference between the structure matrix and the pattern matrix exists only for the two location factors. Factor 1 has negligible correlation with income variables in respect to the unique contributions, thus the correlation with these variables in the structure matrix is high and demonstrates segregation at the expense of common contributions. The unique contributions of factor 4 have low correlations with travel times (the highest one is for the CBD, -0.34), but at the expense of common contributions the correlations are much higher (up to -0.63) in the structure matrix. For each factor we interpolate its score to a raster[1] in order to create a continuous representation of its geographical distribution. These raster maps are presented in Figures 3-6, where factor scores are grouped in nine classes.

Factor 1 is highly positively correlated with travel times to centres and thus represents locations farther from centres, where high income households live as opposed to low income population. In Figure 3 it is represented as a central core of low scores and belts, which demonstrate the fact that high income households prefer to live farther from the central part. Note also that in the north the third and the fourth belts cross the administrative boundaries of Lyon and Villeurbanne. Indeed, this district named Caluire-et-Cuire is urbanised and has metro and trolleybus links with the central part of Lyon.

The spatial distribution of factor 4 is different irrespective of its correlation with factor 1. Factor 4 is highly positively correlated with low income households and highly negatively correlated with high income households. For its common contributions it is also important to be closer to urban centres. Figure 4 clearly demonstrates that the area with the highest scores of factor 4 is located in the central part of Lyon and overlaps with Guillotière – a problematic low income area located remarkably close to the CBD, populated by immigrants and being the object of the specific attention of the police.

Factor 2 and factor 3 account for internal apartment attributes. The former describes big and expensive apartments, the highest concentration of which is seen in the most picturesque locations in Cité International and the west of Croix-Rousse (Figure 5). Factor 3 deals with older apartments (whose attribute is cellars) in bad condition, the maximum is observed in the western part of the 6th arrondissement, while there are areas of low scores in the eastern parts of Lyon and Villeurbanne as well as to the south-east from their boundary (Figure 6).

[1] The Inverse Distance Weighted method is used with 12 neighbours, power 2 and output cell size of 10 metres.

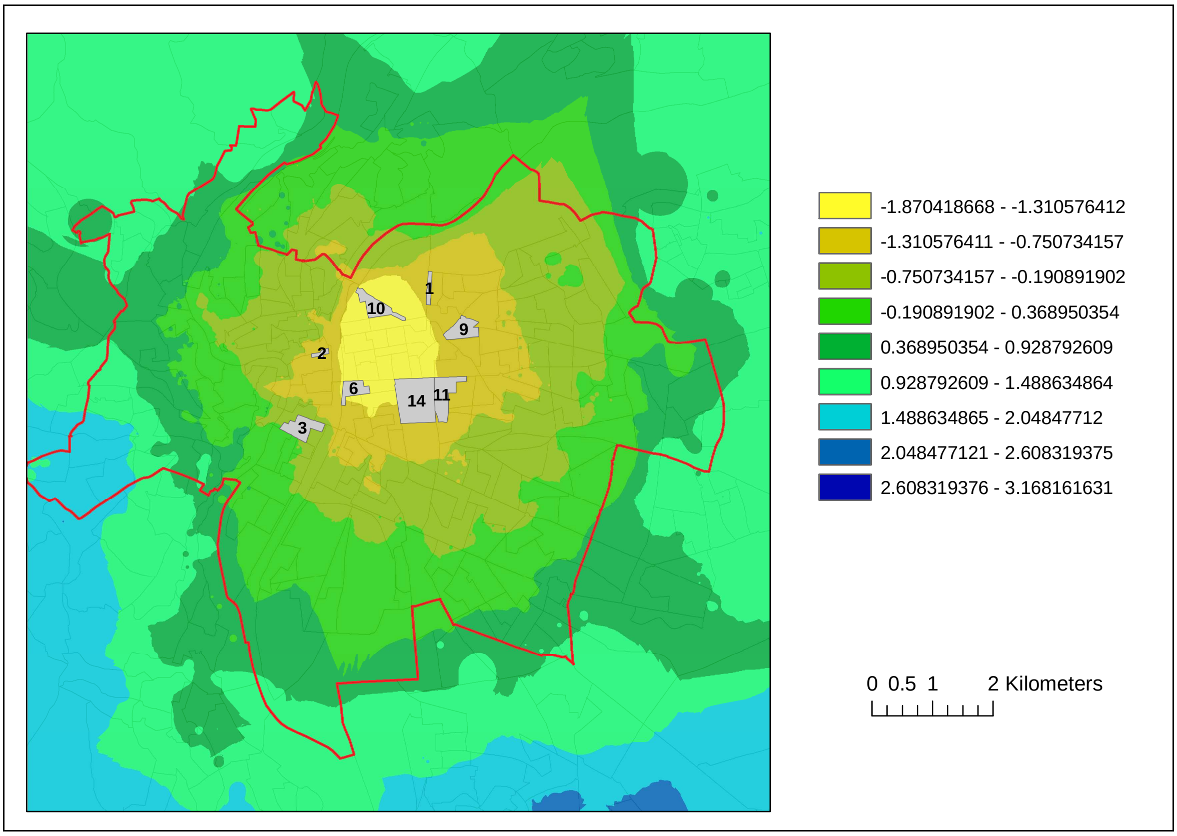

Figure 3. Raster map of factor 1

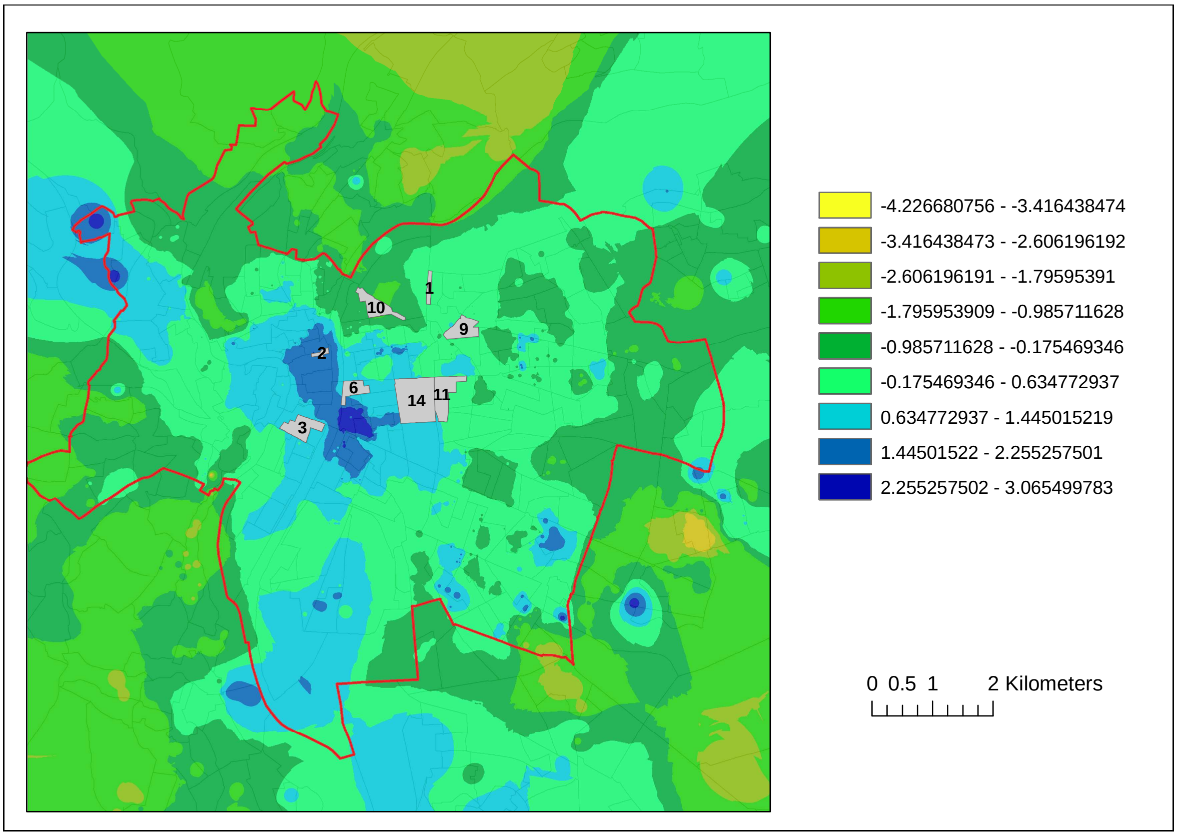

Figure 4. Raster map of factor 4

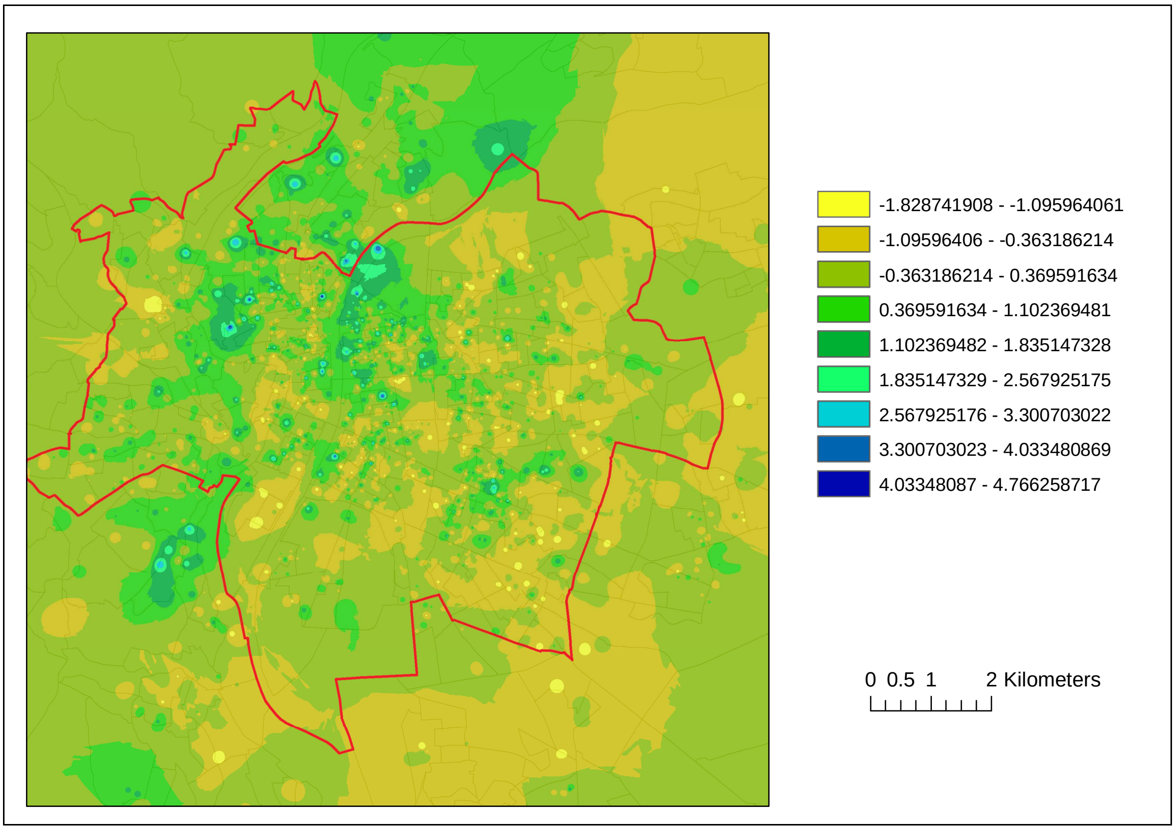

Figure 5. Raster map of factor 2

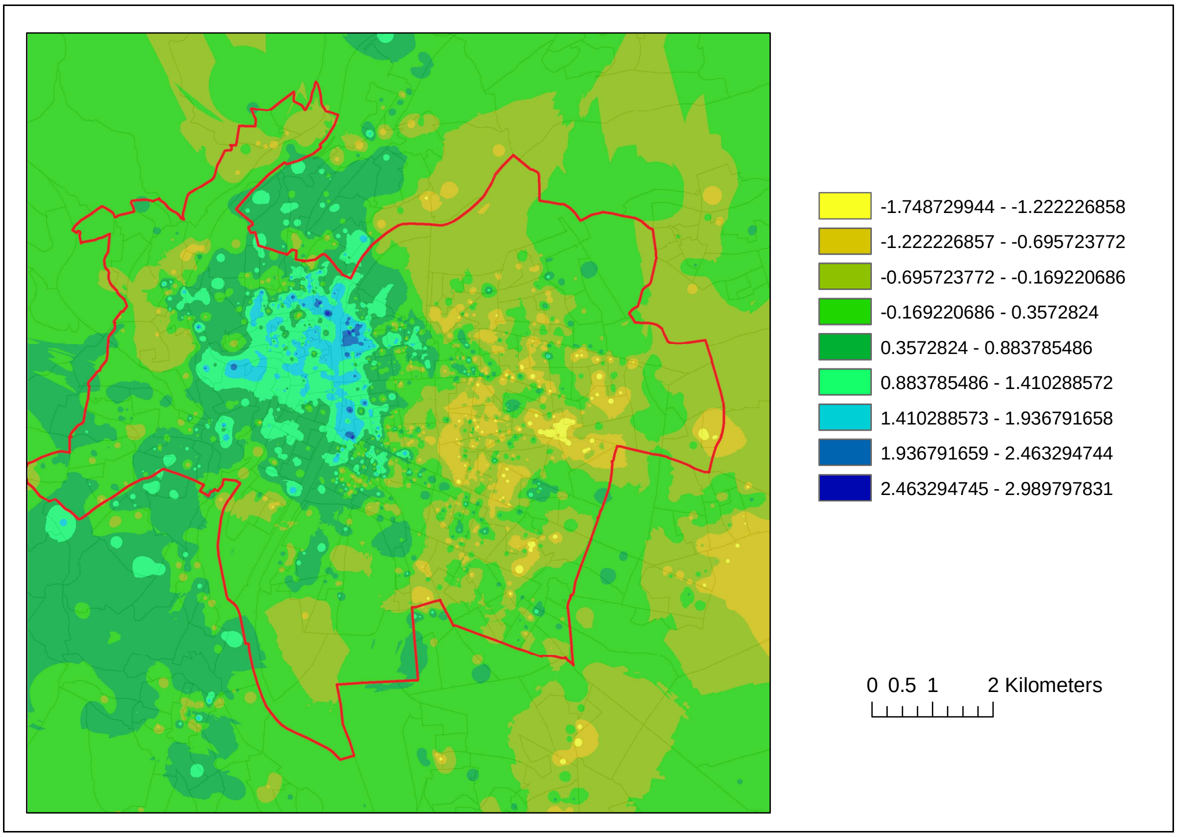

Figure 6. Raster map of factor 3

4. PCA for location attributes

We can execute one more exercise with location attributes by analysing travel times to all fifteen service employment centers. We can do this with PCA. With direct oblimin rotation for fifteen travel time variables and two income variables we obtain three principal components whose eigenvalues are higher than unity. The communalities of variables are very high, ranged from 0.82 to 0.99 with the mean of 0.94. Correlation between the first and the second components is 0.54, between the first and the third is -0.50, while between the second and the third it is -0.32.

The configuration of a raster map of the third principal component is very similar to that of factor 4 presented in Figure 4. Figures 7-8 representing the two first principal components show also the boundary between Lyon and Villeurbanne, which is located to the north-east from Lyon. The first and the second components tell more about the urban structure than factor 1 told. While the spatial distribution of the first component resembles those of factor 1, its core covers bigger areas including eleven of twelve service employment centres located in Lyon (Figure 7). The geographical distribution of the second component is shifted to Villeurbanne, and the three centres of Villeurbanne are located on the diagonal of its central core (Figure 8). Thus, the latent variables highlight the fact that though Lyon and Villeurbanne have many things in common; the influence of their centres is different and not yet amalgamated spatially.

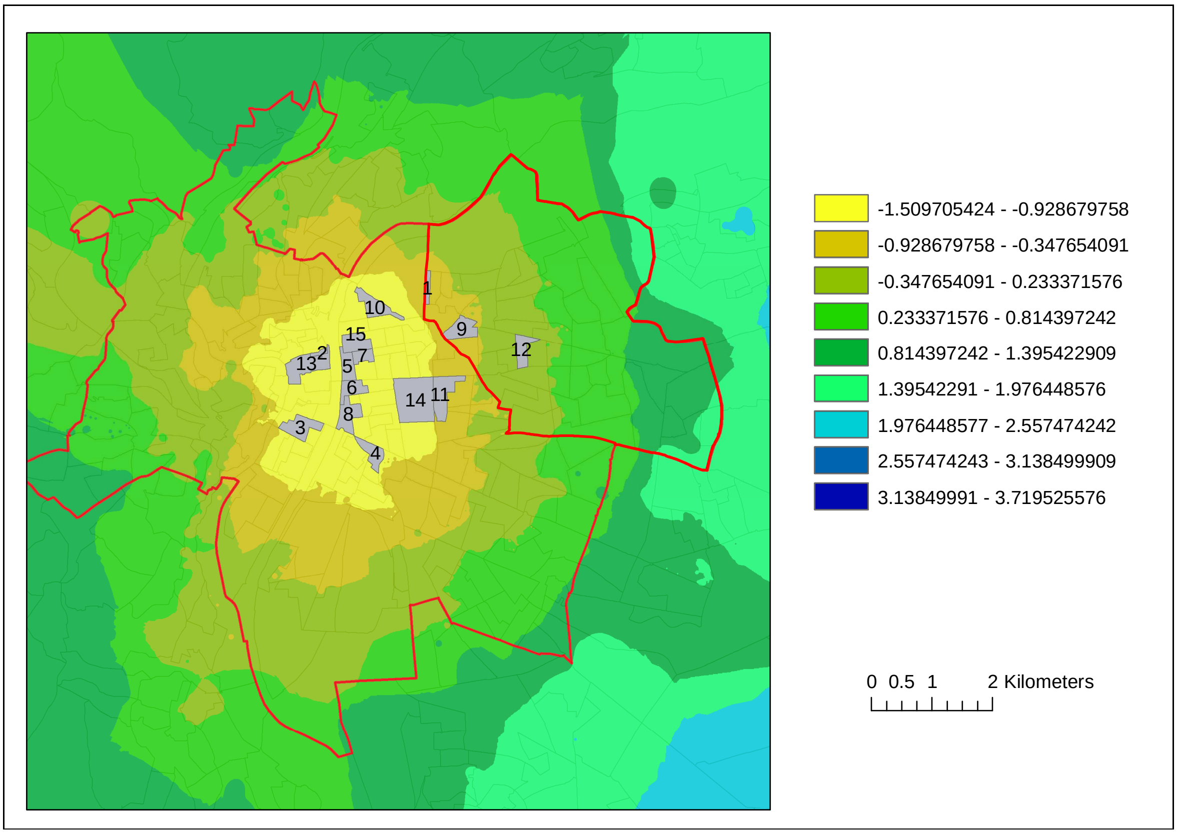

Figure 7. Raster map of the first principal component for location

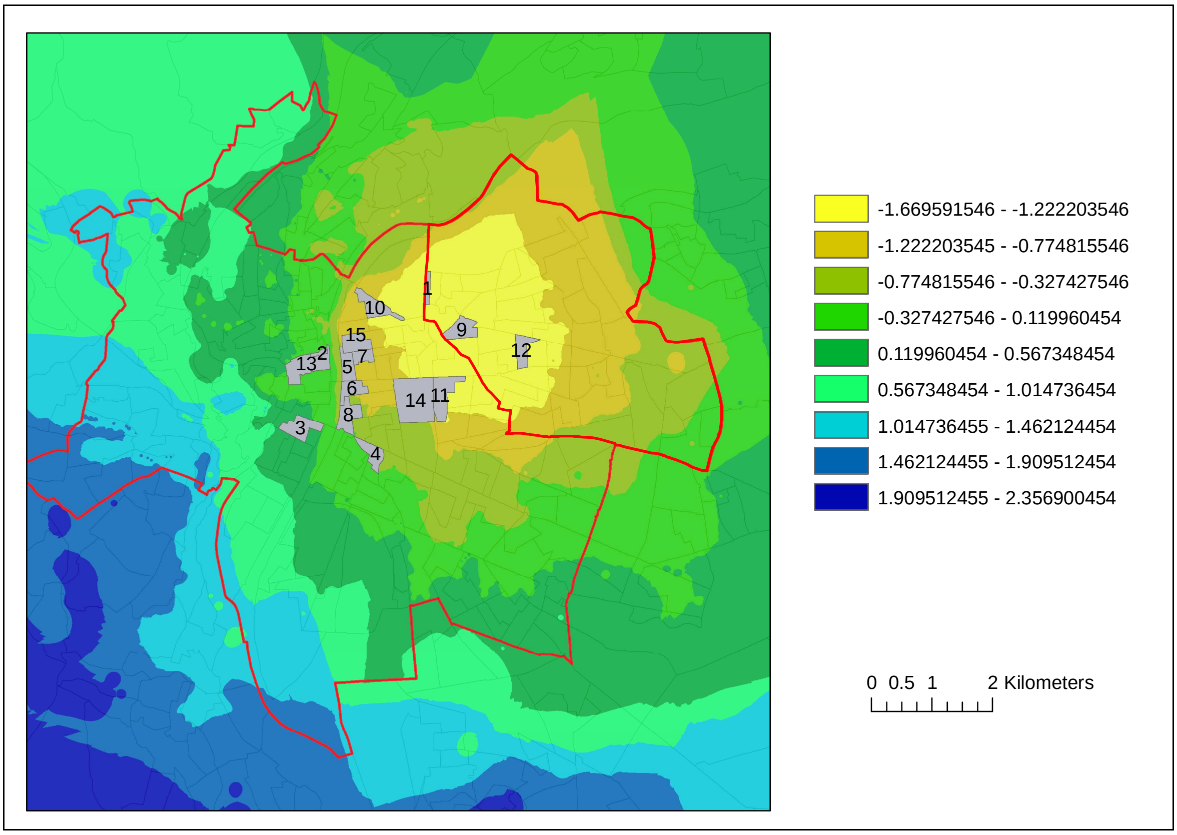

Figure 8. Raster map of the second principal component for location

5. Clusters of principal components

In this section we create the clusters of factor scores with the clustering procedure of K-means, as e.g. in Bourassa *et al*. (1999) and Bourassa *et al*. (2003). As is noted in the latter source, location is the single best criterion to use when defining submarkets. In our study, it is the PCA methodology that allows including all the location attributes. The clusters are created using the three principal components for location reported in the previous section. Figure 9 represents five clusters. Of them, the fourth cluster is a very central location; the third cluster includes some prestigious areas in Lyon, Villeurbanne and Caluire-et-Cuire. The other clusters are more specially dispersed. Probably, it is worth to increase the number of clusters, but this is the subject for a subsequent study, where the clusters can be used as proxies for submarkets.

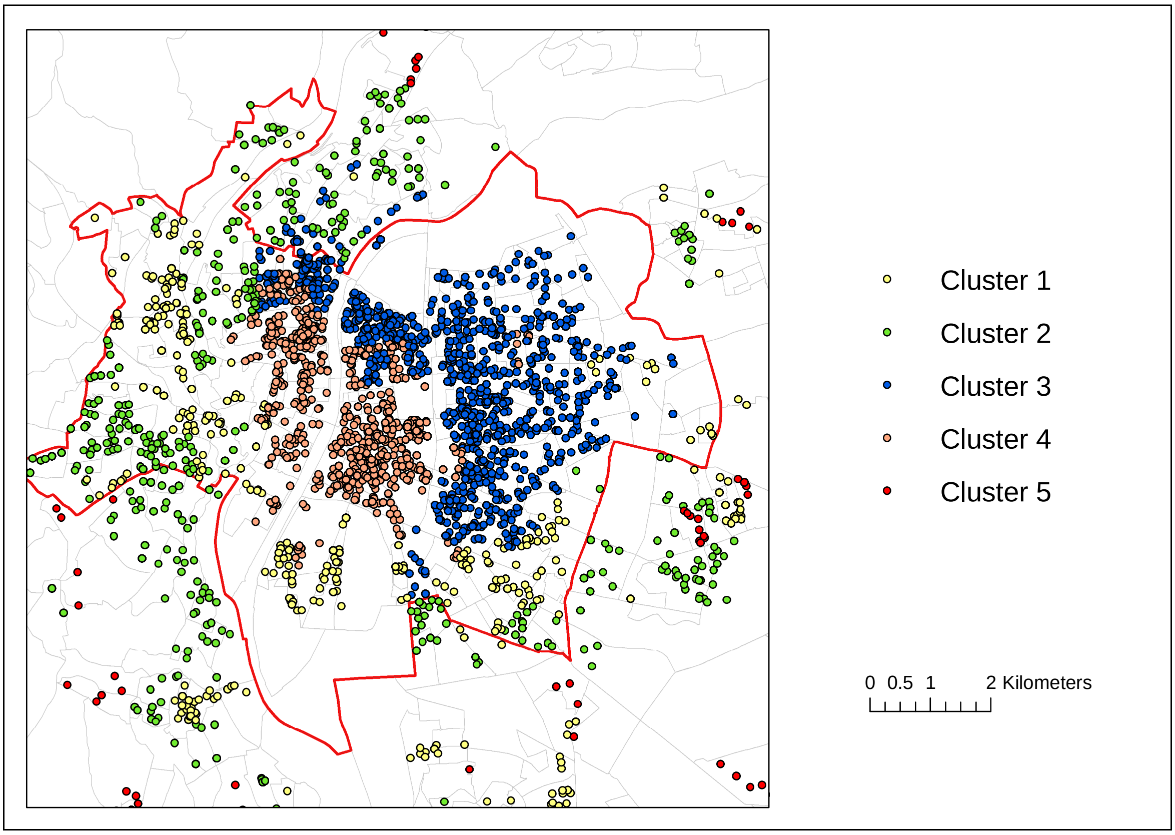

Figure 9. Clusters of location attributes

6. Conclusion

EFA with oblique rotation is found to be applicable for extraction of latent variables providing an insight into apartment attributes and urban structure. The results are intuitively easy to interpret. Factor analysis did not find a strong interaction between apartment attributes and location attributes: separate factors were formed for the two groups. Of the two factors of apartment attributes one accounts for big and expensive apartments and the other represents older apartments in bad conditions. One of the two location factors demonstrate a contradiction with the existing city boundaries in the north, while the other highlights the existence of a problematic low income area in the central part of Lyon similarly to the finding of Des Rosiers *et al*. (2000) in respect to the Quebec Urban Community.

The limitation of EFA is its inability to work with many highly correlated variables. To include into analysis the travel times to all the service employment centres, we applied PCA with non-orthogonal rotation. With more variables included, a more complex latent structure was delineated with separation between the centres of Lyon and those of Villeurbanne.

Thus, both EFA and PCA are found to be useful and illustrative for better understanding the complexity of urban structure. Future study should focus on the clusters of factors and/or principal components as proxies of apartment submarkets.

Acknowledgements

The study is part of project PLAINSUDD sponsored though French ANR (number ANR-08-VD-00). The authors are grateful to Cristina Pronello from Politecnico di Torino for fruitful suggestions on factor analysis and to Nicolas Ovtracht and Valérie Thiebaut for calculation of travel times and the coordinates of apartments.

References

Bonnafous, A. (1973). *La logique de l'investigation économétrique*, Dunod, Paris/Brussels/Montréal.

Bourassa, S. C., Hamelink, F., Hoesli, M., and MacGregor, B. D. (1999). Defining Housing Submarkets, *Journal of Housing Economics*, 8 (2), pp. 160-183.

Bourassa, S. C., Hoesli, M., and Peng, V. S. (2003). Do Housing Submarkets Really Matter? *Journal of Housing Economics*, 12 (1), pp. 12-28.

Breckler, S. J. (1984). Empirical validation of affect, behaviour, and cognition as distinct component of attitude, *Journal of Personality and Social Psychology*, 47 (6), pp. 1191-1205.

Crites, S.L., Jr., Fabrigar, L. R., and Petty, R. E. (1994). Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues, *Personality and Social Psychology Bulletin*, 20 (6), pp. 619-634.

Dale-Johnson, D. (1982). An Alternative Approach to Housing Market Segmentation Using Hedonic Price Data, *Journal of Urban Economics*, 11 (3), pp. 311-332.

Des Rosiers, F., Thériault, M., Villeneuve, P.-Y. (2000). Sorting out access and neighbourhood factors in hedonic price modelling, *Journal of Property Investment and Finance*, 18 (3), pp. 291-315.

Des Rosiers, F. and Thériault, M. (2008). Mass Appraisal, Hedonic Price Modelling and Urban Externalities: Understanding Property Value Shaping Process, in Kauko, T. and d'Amato, M. (eds.), *Mass appraisal methods: an international perspective for property valuers*, Blackwell Publishing Ltd., 332 p.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research, *Psychological Methods*, 4(3), pp. 272-299.

Kolenikov, S. and Angeles, G. (2004). *The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices*. Working paper WP-04-85, MEASURE/Evaluation project, Caroline Population Center, University of North Caroline, Chapel Hill.

Kryvobokov, M. (2009). *Is it worth identifying service employment (sub)centres for modelling apartment prices? The case of Lyon, France*, Presentation paper, ERES conference, Stockholm, 29 p.

Maclennan, D. and Tu, Y. (1996). Economic perspectives on the structure of local housing systems, *Housing Studies*, 11 (3), pp. 387-406.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of Political Economy*, 82, p. 34-55.

Vias, S. and Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis, *Health Policy and Planning*, 21(6), pp. 459-468.

Warner, R. M. (2007). *Applied statistics: From bivariate through multivariate techniques*. SAGE Publications, Thousand Oaks, CA, 1128 p.

West, S. G., Finch, J. F., and Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In Hoyle, R. H. (Ed.), *Structural equation modeling: Concepts, issues and applications*, Newbury Park, CA: Sage, pp. 56-75.