Università degli Studi di Napoli
Federico II

Multidimensional Analysis for the definition of the choice set in Discrete Choice Models

Antonio Lucadamo

Tesi di Dottorato in
Statistica

XIX Ciclo

Multidimensional Analysis for the definition of the choice set in Discrete Choice Models

Napoli. 30 novembre 2006

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Discrete choice methods model a decision-maker's choice among a set of mutually exclusive and collectively exhaustive alternatives. They are used in a variety of disciplines (economics, transportation, psychology, public policy, etc.) in order to inform policy and marketing decisions and to better understand and test hypotheses of behavior. The standard tool for modeling individual choice behavior is the choice model based on the random utility hypothesis. These models have their foundation in classic economic consumer theory, which is the source of many of the important assumptions of these models. Economic consumer theory states that consumers are rational decision makers. So when they are faced with a set of possible consumption bundles of good, they assign preferences to each of the various bundles and then choose the most preferred bundle from the set of affordable alternatives. If we consider the following properties:

- **Completeness** Any two bundles can be compared, i.e. either a is preferred to b, or b is preferred to a or they are equally preferred;

- **Transitivity** If a is preferred to b and b is preferred to c, then a is preferred to c;

- **Continuity** If a is preferred to b and b is arbitrarily close to a then c is preferred to b;

it can be shown that there exists a continuous function, *utility function*, that associates a real number to each possible bundle, such as it summarizes the preference ordering of the consumer (Varian 1992). Consumer behavior can then be expressed as an optimization problem in which the consumer selects the consumption bundle such that their utility is maximized subject to their budget constraint. This optimization function can be solved to obtain the demand function. The demand function can be substituted back into the utility equation to derive indirect utility function, which is the maximum utility that is achievable under the given prices and income. The indirect utility function is what generally is used in discrete choice models, and in the rest of the dissertation it will be indicated simply as *utility*.

There are several extensions to the classical consumer theory that are important to discrete choice models. In fact consumer theory assumes homogeneous goods and therefore the utility is a function of quantities only and not attributes. A first extension is then, the fact that the attributes of the goods that determine the utility they provide and therefore utility can be expressed as a function of the attributes of the commodities (Lancaster 1966).

Second is the concept of random utility theory (Thurstone 1927), (Marschak 1960); according to this theory, differently by consumer theory which assumes deterministic behavior, the individual choice behavior is intrinsically probabilistic. The idea behind this theory is that while the decision maker may have perfect discrimination capability, the analyst has incomplete information and therefore uncertainty must be taken into account. Therefore utility is modeled as a random variable, consisting of an observable and an unobservable component.

Finally consumer theory deals with continuous products. Calculus is used to derive many of the key results, and so a continuous space of alternatives is required. Discrete choice theory deals instead with a choice among a set

of finite, mutually exclusive alternatives and so different techniques need to be used. However the underlying hypotheses of random utility remain intact. The standard technique for modeling individual choice behavior is then the discrete choice model derived from random utility theory. The model is based on the notion that the individual derives utility by buying or choosing an alternative. Usually the models assume that the individual selects the alternative that has the maximum utility, but other decision protocols can be used. The utilities are latent variables, and the actual choice, which is what can be observed, is a manifestation of the underlying utilities. The utilities are assumed to be a function of the attributes of the alternatives and the characteristics of the decision maker, that are introduced to capture heterogeneity across individuals. The final component of the utility is a random disturbance term. Assumption on the distribution of the disturbances lead to various choice models (probit and logit for example). The outputs of the models are the probabilities of an individual selecting each alternative. These individual probabilities can then be aggregated to produce forecasts for the population. In discrete choice models simplifying assumptions are made, in order to maintain a parsimonious and tractable structure.

1.1 Behavioral Theory and Discrete Choice Models

In the behavioral science and economic communities there has been much debate on the validity of discrete choice models, due to the strong assumptions and simplifications that are made. A large gap between behavioral theory and discrete choice analysis exists, due to the driving forces behind the two disciplines; in fact while discrete choice modelers are focused on mapping inputs to the decision, behavioral researchers aim to understand the nature of how decisions come about, or the decision-process itself. To understand the difference it's possible to consider the figure 1.1

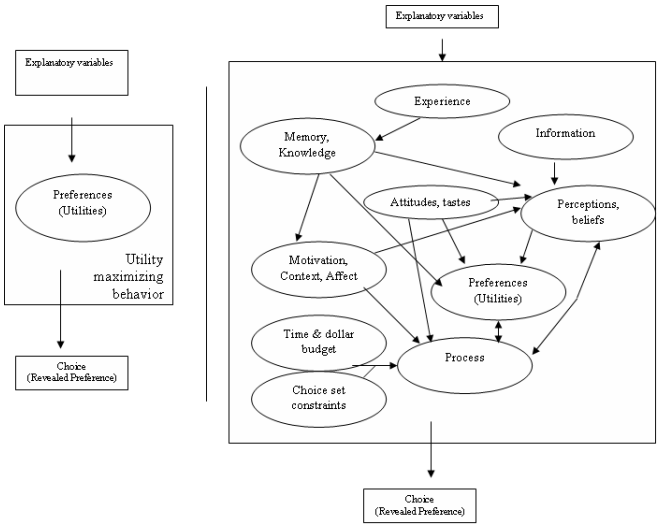

Figure 1.1: *The gap between Discrete Choice Models(left) and the complexity of Behavior*

It's clear that preferences are unobservable, but they are assumed to be a function of explanatory variables as well as unknown parameters and a disturbance term. The choice is a manifestation of preferences and the typical assumption is that the alternative with the maximum utility is chosen. This model is often described as an optimizing "black box" (Ben-Akiva, Fadden, Garling, Gopinath, Walker, Bolduc, Borsch-Supan, Delquié, Larichev, Morikawa, Polydoroulou & Rao 1999), because the model directly links the observed inputs to the observed output and, thereby assumes that the model implicitly captures the behavioral choice process. If we look at the right side of the figure we can see that in the reality there are many unobserved factors that influence the choice and so it's important to understand if the discrete choice models are an adequate representation of the reality. Multinomial Logit is the standard model used in these circumstances, but sometimes, different specifications, that we will see in the dissertation, are necessary. In this sense, to reduce the gap between real behavior and discrete choice analysis,

the findings and the techniques from related fields have been very important. Psychometricians, i. e. in their quest to understand behavioral constructs, have pioneered the use of psychometric data, for example, answers to direct survey questions regarding attitudes, perceptions, motivations, affect, etc. A general approach to synthesizing models with latent variables has been advanced by different researchers (Bentler 1980) who developed the structural and measurement equation framework and methodology for specifying and estimating latent variable models. Market researchers instead have long used stated preference data to provide insight on preferences. The basic idea is to obtain a rich form of data on behavior by studying the choice process under hypothetical scenarios designed by the researcher (Luce & Tukey 1964). There are many advantages to these data including the ability to capture responses to products not yet on the market, design explanatory variables such that they are not collinear and have wide variability, control the choice set, etc. However they have also some drawbacks as the fact that they may be not congruent with actual behavior. For this reason, techniques to combine stated and revealed preferences which draw on the relative advantages of each type of data are becoming increasingly popular (Ben-Akiva & Morikawa 1990). Another area of enhancements to discrete choice models is related to the idea that there is heterogeneity in behavior across individuals, and ignoring this heterogeneity can result in forecasting errors. The most straightforward way to address this issue is to capture the so-called "observed heterogeneity" by introducing socio-economic and demographic characteristics in the observed part of the utility function. But some other techniques aimed to capture also unobserved heterogeneity. Another technique is latent class models, which can be used to capture unobservable segmentation regarding tastes, choice set and decision protocols.

1.2 Residential choice models

One of the field in which discrete choice models are applied is the choice of the residence.

The home is where people typically spend most of their time, a common venue for social contact and, for most people, a major financial and personal investment. One's choice of residence also reflects one's choice of surrounding neighborhood, which has a significant impact on one's well-being and quality of life. The topic of residential location choice has, therefore, been of interest to sociologists, psychologists, urban economists, geographers and transportation planners. In the past years there were many studies on this subject, including the relationship between life quality and location, market differentiation in housing demand, social value of urban amenities and neighborhood quality, and effects of spatial policies.

For urban and transportation planning, the concern for the causes and consequences of individual's choice of residence arises from the recognition that it is the values, decisions and actions of the people who are attracted to certain types of land use patterns that ultimately shape the transportation, land-use and urban form. The decision of residential location not only determines the connection between the households with the rest of the urban environment, but also influences the household's activity time budgets and perceived well being. The need for understanding land use-transport linkage at the indivdual level and the debate over whether the influence of urban form is entirely due to individuals placing themselves into residential neighborhoods that support their travel properties points to the need for better models of residential location preferences.

In the past years there has been considerable development in the mathematical modeling of residential activities; based on the trade-off theory, Alonso (Alonso 1964) was the first one to consider the residential location choice based on the concept of utility maximization. The level of utility a households experiences depends on the expenditure in good, size of the land lots,

and distance from the city center. The most criticized aspects of these urban economic studies are:

- The models treat location as a one-dimensional variable and are therefore incapable of handling the common situations of dispersed employment centers and asymmetric development patterns;
- All members of any one socio-demographic class are considered to have identical behavior, which is certainly an oversimplification of reality;
- By reducing the complexity of the housing commodity, which is multidimensional and heterogeneous, to the one-dimensional measure of price, one assumes that many of the important and interesting housing market phenomenon are irrelevant.

These problems was faced with the introduction of the discrete choice analysis (McFadden 1974). In this way it was possible for the analysts to examine the choice behavior based on both accepted and rejected alternatives and to relate spatial behavior to locational characteristics s well as the complex attitudes, preferences and tastes of individuals. The modeling results can thus help devise urban policies that effectively target specific population groups. For these reasons, discrete choice analysis dominates spatial choice theory, even though it was originally developed for non-spatial context such as the choice of transportation mode. Anyway, the spatial characteristics, often create problems for the use of classical discrete choice models, that we will face during the dissertation.

1.3 Outline of the dissertation

The dissertation is organized as follows:

- **Chapter 2** focuses the attention on the Random Utility Model and the derivation of the different discrete choice models. We will show

the properties and characteristics of the most used methods: Binary Logit, Multinomial Logit, Nested and Cross-Nested Logit. At the end of the chapter we show how all these models are derivable considering a general model, the Generalized Extreme Value Models.

- in **Chapter 3** we describe what kind of problems can arise when we consider a discrete choice model in which the alternative to be chosen has spatial implication. We introduce some contiguity measurement that will be useful for the proceeding of the dissertation and then we consider some modifications to the classical logit models that allow to introduce spatial component in the analysis (Spatial Multinomial Logit, Mixture of Logit models, etc.). However, these models don't consider that in some kind of analysis, i.e. destination and residential choice analysis, the number of alternatives is huge and in this case a computational burden for the estimation could be. To solve this problem technique to aggregate alternatives or to sample them and apply the model on a reduced choice-set are introduced, underlining advantages and drawbacks.

- **Chapter 4** intends to show how multidimensional analysis could be a tool to solve some problems related to the spatial dimension and to the size of the choice set. We introduce briefly Principal Component Analysis and therefore we describe the Constrained Principal Component Analysis (CPCA), showing that considering a particular matrix rather than the scalar product matrix, it's possible to carry out a new method to aggregate the alternatives. The CPCA will be showed to be useful also to propose an innovative way to conduce a stratified sampling.

- in **Chapter 5** we shows the usefulness of the two techniques applying them on a data-set relative to the choice of residential location in Zurich area. First of all we build a model on the full choice-set of alternatives and we estimate the parameters; afterwards, we carry out the aggrega-

tion and the sampling of the possible chooses, following the approach we introduce previously; furthermore a simple random sampling has been applied and then the model built for the full choice set has been applied for the samples obtained in the different way. At the end of the chapter we compare the results obtained with the different techniques showing the improvements that we can have with the innovative methodology. The analysis has been carried out combining the use of S-Plus, in which we wrote the code to implement the Multidimensional Analysis, and of BIOGEME (Bierlaire 2003), (Bierlaire 2005), a software that allows the estimation of different Generalized Extreme Value models.

Chapter 2

Discrete Choice Analysis

2.1 Introduction

Discrete choice models are methods used to model a decision-maker's choice among a set of mutually exclusive and collectively exhaustive alternatives. They are usually derived under an assumption of utility-maximizing behavior by the decision maker. The original concepts were developed by Thurstone (Thurstone 1927) in terms of psychological stimuli. Marschak (Marschak 1960) interpreted the stimuli as utility and provided a derivation from utility maximization. Models that can be derived in this way are called random utility models (RUM). In these models, a decision maker, labeled n, faces a choice among J alternatives. The decision maker would obtain a certain level of utility from each alternative. The utility that he obtains from alternative j is $U_{nj},\ j = 1, ..., J$. The individual is always assumed to select the alternative with the highest utility, but the analyst doesn't know the utilities with certainty and so they are treated as random variables. From this perspective the choice probability of alternative i is equal to the probability that the utility of alternative i, U_{in}, is greater than the

utilities of all others alternatives in the choice set:

$$P(i|C_n) = Pr[U_{in} \geq U_{jn}, \quad allj \in C_n] \tag{2.1}$$

In this approach a joint probability distribution is assumed for the set of random utilities $\{U_{in}, i \in C_n\}$

2.2 Random Utility

Manski (Manski 1973) identified four distinct sources of randomness:

- unobserved attributes;
- unobserved taste variation;
- measurement errors and imperfect information;
- instrumental variables.

Unobserved attributes The vector of attributes affecting the decision is incomplete, so the utility function

$$U_{in} = U(z_{in}, S_n, z_{in}^u) \tag{2.2}$$

includes an element z_{in}^u which is a random variable and consequently the utility is in itself random.

Unobserved taste variations The utility function

$$U_{in} = U(z_{in}, S_n, S_n^u) \tag{2.3}$$

may have an unobserved argument S_n^u which varies among individuals. Also in this case U_{in} is a random variable.

Measurement errors The true utility function is

$$U_{in} = U(\tilde{z}_{in}, S_n) \tag{2.4}$$

In this case we can only observe z_{in} which is an imperfect measurement of $\tilde{z}_{in}$. We can substitute

$$\tilde{z}_{in} = z_{in} + \tilde{\varepsilon}_{in} \tag{2.5}$$

where $\tilde{\varepsilon}_{in}$ is the unknown measurement error, into the utility function to get the new utility:

$$U_{in} = U(z_{in} + \tilde{\varepsilon}_{in}, S_n) \tag{2.6}$$

which contains a random element.

Instrumental Variables In this case the true utility function is

$$U_{in} = U(\tilde{\tilde{z}}_{in}, S_n) \tag{2.7}$$

and some elements of $\tilde{\tilde{z}}_{in}$ are not observable. Therefore we substitute

$$\tilde{\tilde{z}}_{in} = g(z_{in}) + \tilde{\tilde{\varepsilon}}_{in} \tag{2.8}$$

into the utility function to obtain:

$$U_{in} = U[g(z_{in}) + \tilde{\tilde{\varepsilon}}_{in}, S_n] \tag{2.9}$$

where g denotes the imperfect relationship between instruments and attributes and $\tilde{\tilde{\varepsilon}}_{in}$ is again a random error. Therefore in general, we can express the random utility of an alternative as a sum of observable and unobservable components of the total utilities, in the following

way:

$$U_{in} = V(z_{in}, S_n) + \varepsilon(z_{in}, S_n) = V_{in} + \varepsilon_{in} \tag{2.10}$$

where V_{in} is the systematic utility and ε_{in} is the random disturbance. The systematic utility V_{in} is really the expected value of the perceived utility between all the individuals who have the same choice set of individual i. The random disturbance ε_{in} is the difference, from this expected value, of the utility perceived from decisor i. So we have: $V_{in} = E[U_{in}]$ and $\sigma_{in}^2 = Var[U_{in}]$. Furthermore we know that V_{in} is a deterministic value with the following mean and variance:
$E[V_{in}] = V_{in}$ and $Var[V_{in}] = 0$
then we can calculate the expected value and the variance of the disturbance:
$E[\varepsilon_{in} = 0]$ and $Var[\varepsilon_{in} = \sigma_{in}^2]$
Furthermore generally V_{in} is supposed to be linear in parameters so we can express it in the following way:

$$V_{in} = \beta' x_{in} \tag{2.11}$$

where β are parameters to be estimated and x are explicative variables. The probability to select the alternative i from the choice set C_n becomes:

$$P(i|C_n) = Pr[V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \quad allj \in C_n] \tag{2.12}$$

We can derive a specific random utility model if we have an assumption about the joint probability distribution of the full set of disturbances $\{\varepsilon_{jn}, j \in C_n\}$. One logical assumption is that the disturbances are normal distributed. In this case we have the probit model, but it has the disadvantage of not having a closed form and in this circumstance

we have to express the choice probability as an integral. For this reason the assumption that the disturbances are independent and identically Gumbel distributed is made. In this way we obtain a logit model. If we have only two alternatives we obtain a Binary Logit Model and in this case the assumption is simply that the difference $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ is logistically distributed that is equivalent to the assumption of Gumbel distribution if we have more than two alternatives. In this second case we have a Multinomial Logit Model. Before describing these models we must introduce the Gumbel distribution with its properties that will be useful in the following sections.

2.3 Gumbel Distribution

If ε is Gumbel distributed then

$$F(\varepsilon) = exp[-e^{-\mu(\varepsilon-\eta)}], \quad \mu > 0 \tag{2.13}$$

and

$$f(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} exp[-e^{-\mu(\varepsilon-\eta)}] \tag{2.14}$$

where η is a location parameter and μ is a positive scale parameter, the distribution has the following properties:

- The mode is η;
- The mean is $\eta + \frac{\gamma}{\mu}$ where γ is Euler constant ($\sim 0,577$);
- The variance is $\frac{\pi^2}{6\mu^2}$;
- If we consider a scalar constant $\alpha > 0$ then also $\alpha\varepsilon + V$ is Gumbel distributed with parameters $(\alpha\eta + V, \frac{\mu}{\alpha})$

- If ε_1 and ε_2 are independent Gumbel-distributed variates with parameters $(\eta_1, \mu) and (\eta_2, \mu)$, respectively, then $\varepsilon^* = \varepsilon_1 - \varepsilon_2$ is logistically distributed:

$$F(\varepsilon^*) = \frac{1}{1 + e^{\mu(\eta_2 - \eta_1 - \varepsilon^*)}} \tag{2.15}$$

- If ε_1 and ε_2 are independent Gumbel-distributed with parameters (η_1, μ) and (η_2, μ), respectively, then $max(\varepsilon_1, \varepsilon_2)$ is Gumbel distributed with parameters
  $(\frac{1}{\mu} ln(e^{\mu\eta_1} + e^{\mu\eta_2}), \mu)$
- If $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j)$ are J independent Gumbel distributed variables with parameters (η_1, μ), (η_2, μ) ,...,(η_j, μ) then $max(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j)$ is Gumbel distributed with parameters
  $\left(\frac{1}{\mu} ln \sum_{j=1}^{J} e^{\mu\eta_j}, \mu\right)$

This last property is very important for our purpose.

2.4 Binary Logit

The binary logit model arises, as we said, from the assumption that $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ is logistically distributed:

$$F(\varepsilon_n) = \frac{1}{1 + e^{-\mu\varepsilon_n}}, \qquad \mu > 0, \text{-}\infty < \varepsilon_n < \infty \tag{2.16}$$

$$f(\varepsilon_n) = \frac{\mu e^{-\mu\varepsilon_n}}{(1 + e^{-\mu\varepsilon_n})^2} \tag{2.17}$$

where μ is a positive scale parameter. So under the assumption that ε_n is logistically distributed, the choice probability for alternative i is

given by:

$$\begin{aligned} P_n(i) &= Pr(U_{in} \geq U_{jn}) \\ &= \frac{1}{1 + e^{\mu(V_{in} - V_{jn})}} \\ &= \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}} \end{aligned} \tag{2.18}$$

If V_{in} and V_{jn} are linear in their parameters, we have:

$$\begin{aligned} P_n(i) &= \frac{e^{\mu \beta' x_{in}}}{e^{\mu \beta' x_{in}} + e^{\mu \beta' x_{jn}}} \\ &= \frac{1}{1 + e^{-\mu \beta' (x_{in} - x_{jn})}}. \end{aligned} \tag{2.19}$$

In this circumstances we cannot distinguish the parameter μ from the overall scale of the β' and for convenience we can do the arbitrary assumption that $\mu = 1$. In this case, according to Gumbel distribution the variances of ε_{in} and ε_{jn} are both $\frac{\pi^2}{6}$ implying that the variance of $\varepsilon_{jn} - \varepsilon_{in} = \frac{\pi^2}{3}$.

2.4.1 Estimation of Binary logit with Maximum Likelihood technique

In the binary logit we have that:

$$y_{in} = \begin{cases} 1, \, if \; n \; chose \; i \\ 0, \, if \; n \; chose \; j \end{cases} \tag{2.20}$$

Furthermore we have two vectors of attributes x_{in} and x_{jn} each containing values of the K esplicative variables. So, given a sample of N observations, the problem is to find estimates of $\hat{\beta}_1$,$\hat{\beta}_2$,...,$\hat{\beta}_k$. We can consider the likelihood of any sample of N observations, the likelihood

of the entire sample is the product of the likelihood of the individual observations, because they are by assumption drawn at random from the whole population. So we have:

$$\ell * (\beta_1, \beta_2, ..., \beta_k) = \prod_{n=1}^{N} P_n(i)^{y_{in}} P_n(j)^{y_j n} \tag{2.21}$$

where $P_n(i)$ is a function of $\beta_1, ..., \beta_k$. In general it's better to use the logarithm of this function:

$$\ell(\beta_1, ..., \beta_k) = \sum_{n=1}^{N} [y_{in} \log P_n(i) + y_{jn} \log P_n(j)] \tag{2.22}$$

but we know that $y_{jn} = 1 - y_{in}$ and $P_n(j) = P_n(i)$ so we can rewrite:

$$\ell(\beta_1, ..., \beta_k) = \sum_{n=1}^{N} \{y_{in} \log P_n(i) + (1 - y_{in}) \log[1 - P_n(i)] \tag{2.23}$$

Now we have to find the maximum of ℓ differentiating the equation respect each of the $\beta's$ and setting the partial derivatives equal to zero:

$$\frac{\partial \ell}{\partial \hat{\beta}_k} = \sum_{n=1}^{N} \left\{ y_{in} \frac{\partial P_n(i) / \partial \hat{\beta}_k}{P_n(i)} + y_{jn} \frac{\partial P_n(j) / \partial \hat{\beta}_k}{P_n(j)} \right\} = 0, \quad k = 1, ..., K \tag{2.24}$$

The maximum likelihood estimates are consistent, asymptotically efficient and asymptotically normal. The asymptotic variance-covariance matrix is given by:

$$- \Im[\nabla^2 \ell]^{-1} \tag{2.25}$$

where$\nabla^2\ell^{-1}$ is the matrix of second derivatives of the log likelihood function. So the entry in the kth row and the lth column is:

$$[\nabla^2\ell]_{kl} = \frac{\partial^2\ell}{\partial\beta_k\partial\beta_l} \tag{2.26}$$

The problem is that we don't know the actual values of the parameters or the distribution of x_{in} and $x_{jn}$and so we generally use an estimated variance-covariance matrix and the sample distribution of x_{in} and x_{jn} to estimate their distribution. Thus we use:

$$\Im\left[\frac{\partial^2\ell}{\partial\beta_k\partial\beta_l}\right] \cong \sum_{n=1}^{N}\left[\frac{\partial^2\left[y_{in}\log P_n(i) + y_{jn}\log P_n(j)\right]}{\partial\beta_k\partial\beta_l}\right]_{\beta=\hat{\beta}} \tag{2.27}$$

In many occasions the equations that come from 2.24 are non-linear and so there is a computational problem. If the second derivatives can be computed without great difficulty and the likelihood function is globally concave, we can use the Newton-Rapshon method. It's composed of 4 steps:

- **Step1**: First of all we have to choose an initial arbitrary value for $\hat{\beta}_0 = [\beta_{01}, \beta_{02}, ..., \beta_{0K}]'$. Furthermore we must introduce an iteration counter and set it in the following way: $\omega = 0$ and we have also to set e_1 and e_2 to be small positive number.

- **Step2**: The second step consists to linearize the function $\nabla\ell(\beta)$ around $\hat{\beta}_\omega$. The approximate first-order conditions are given by: $\nabla\ell(\hat{\beta}_\omega) + \nabla^2\ell(\hat{\beta}_\omega)(\hat{\beta} - \hat{\beta}_\omega) = 0$

- **Step3**: In the third step we have to solve the linearized form for $\hat{\beta}_{\omega+1} = \hat{\beta}_\omega - [\nabla^2\ell(\hat{\beta}_\omega)]^{-1}\nabla\ell(\hat{\beta}_\omega)$.

- **Step4**: In this step we must check if $\hat{\beta}_{\omega+1} - \hat{\beta}_\omega$ is small. This

happens if:

$$\left[\frac{1}{K}\sum_{k=1}^{K}(\hat{\beta}_{\omega+1,k}-\beta)^2\right]^{1/2} < e_1$$

and

$$\left|\frac{\hat{\beta}_{\omega+1,k}-\hat{\beta}}{\hat{\beta}}\right| < e_2$$

If the condition in this step are satisfied we terminate with $\hat{\beta}_{\omega+1}$ as the solution. Otherwise we must set $\omega = \omega + 1$ and come back to step 2.

There are also others method to estimate the parameters, all of them finding a direction where the log likelihood function is increasing and then searching along that direction for the best possible estimate.

2.5 Multinomial Logit

As we showed, the probability that any element i in C_n is chosen by the decision maker n can be expressed according 2.12. We can rewrite it in the following way:

$$P(i|C_n) = Pr(\varepsilon_{jn} \leq V_{in} - V_{jn} + \varepsilon_{in}, \quad \forall j \in C_n, \quad j \neq i)$$

(2.29)

Any particular multinomial choice model can be derived using the previous equation given specific assumptions on the joint distribution of the disturbances. There are different way to derive $P_n(i)$. The most

insightful is to reduce the multinomial choice problem to a binary one. To do this we can note that

$$U_{in} \geq U_{jn}, \quad \forall j \in C_n, \quad j \neq i, \tag{2.30}$$

is equivalent to

$$U_{in} \geq maxU_{jn}, \quad \forall j \in C_n, \quad j \neq i \tag{2.31}$$

In this way we create a "composite" alternative and the utility of the best alternative $j \neq i$ represent the utility of the entire composite. So we have this situation:

$$P_n(i) = Pr[V_{in} + \varepsilon_{in} \geq max(V_{jn} + \varepsilon_{jn})] \tag{2.32}$$

To calculate this probability we have to derive the distribution of the utility of the composite alternative from the underlying distribution of the disturbances. We can show as in Domencich and McFadden (Domencich & McFadden 1975) that if the disturbances are

- indipendently distributed;
- identically distributed;
- Gumbel distributed with a location parameter η and a scale parameter $\mu \geq 0$;

then the probability that alternative i will be chosen is

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} \tag{2.33}$$

2.5.1 Derivation of Multinomial Logit

If we assume that $\eta = 0$ for all the disturbances and if we order the alternatives so that $i = 1$, then, as showed by Ben-Akiva and Lerman (Ben-Akiva & Lerman 1985), we have:

$$P_n(1) = Pr[V_{1n} + \varepsilon_{1n} \geq \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn})] \tag{2.34}$$

We can define

$$U_n^* = \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn}) \tag{2.35}$$

that from the last property showed in section 2.3 is Gumbel distributed with parameters $(\frac{1}{\mu} ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}, \mu)$. We can also write
$U_n^* = V_n^* + \varepsilon_n^*$ where
$V_n^* = \frac{1}{\mu} ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}$ and ε_n^* is Gumbel distributed with parameters $(0, \mu)$. So the 2.34 become:

$$\begin{aligned} P_n(1) &= Pr[V_{1n} + \varepsilon_{1n} \geq \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn})] \\ &= Pr[(V_n^* + \varepsilon_n^*) - (V_{1n} + \varepsilon_{1n}) \leq 0], \end{aligned} \tag{2.36}$$

and by one other of the previous properties we have

$$\begin{aligned} P_n(1) &= \frac{1}{1 + e^{\mu(V_n^* - V_{1n})}} \\ &= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + e^{\mu V_n^*}} \\ &= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + exp(ln \sum_{j=2}^{J_n} e^{\mu V_{jn}})} \\ &= \frac{e^{\mu V_{1n}}}{\sum_{j=1}^{J_n} e^{\mu V_{jn}}} \end{aligned} \tag{2.37}$$

In the last equation there is the presence of the scale parameter μ. It's not identifiable, but generally it's used to set it to an arbitrary value, such as 1.

2.6 Properties of the logit

One of the most discussed aspects of the multinomial logit is the Independence from Irrelevant Alternatives property (IIA). This property states that for any two alternatives i and k the ratio of the logit probabilities

$$\begin{aligned} \frac{P_n(i)}{P_n(k)} &= \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} \\ &= \frac{e^{V_{ni}}}{e^{V_{nk}}} \\ &= e^{V_{ni} - V_{nk}} \end{aligned} \tag{2.38}$$

does not depend on any alternatives other than i and k. So the relative odds of chosing i over k are the same no matter what other alternatives are available or what the attributes of the other alternatives are. Since the ratio is independent from alternatives other than i and k, it is said to be independent from irrelevant alternatives. This assumption is realistic in some choice situation, but sometimes it can be clearly inappropriate. One of the classical example is the famous red-bus blue-bus problem. In this problem there is a traveler who has to choice if going to work by car or taking a blue bus. For simplicity we can assume that the representative utilities of the two modes are the same, so the probabilities are equal: $P_c = P_{bb} = \frac{1}{2}$ where c is the car and bb the blue bus. The ratio is $P_c/P_{bb} = 1$. If we suppose that a red bus is introduced, probably the traveler considers the red bus to be exactly like the blue bus, so the ratio of this probabilities is one: $P_{rb}/P_{bb} =$

1. However in the logit model the old ratio doesn't change whether or not this other alternative exists. The only probabilities for which $P_c/P_{bb} = 1$ and $P_{rb}/P_{bb} = 1$ are $P_c = P_{bb} = P_{rb} = 1/3$, which are the probabilities that the logit model predicts. In real life however we would expect the probability of taking a car to remain the same when a new bus is introduced that is exactly the same as the old bus. We would also expect the original probability of taking bus to be split between the two buses after the second one is introduced. That is we would expect $P_c = 1/2$ and $P_{bb} = P_{rb} = 1/4$. In this case the multinomial logit model overestimates the probability of taking either of the buses and underestimates the probability of taking a car and so is not appropriate in this case. In this situation we must search for a better model specification:

- find alternatives with missing or mis-specified variables
- point toward an acceptable nested logit structure that we will see in next sections

To verify if the IIA property holds many test exist (McFadden, Tye & Train 1977). We can divide them in two groups:

- Estimate a model with a subset of the choice set. Reject IIA if the parameter estimates differ from the full choice set estimates.
  - Hausman and McFadden
  - McFadden,Tye and Train
  - Small-Hsiao (Small & Hsiao 1982)
- Implement a Lagrange multiplier test of IIA with the full set of alternatives
  - McFadden test

2.6.1 Estimating model with choice subsets

If we suppose IIA holds. Then:

$$P(i|C_n) = \frac{exp(\mu\beta' x_{in})}{\sum_{j\in C_n} exp(\mu\beta' x_{jn})} \tag{2.39}$$

and

$$P(i|\tilde{C}_n \subseteq C_n) = \frac{exp(\mu\beta' x_{in})}{\sum_{j\in \tilde{C}_n} exp(\mu\beta' x_{jn})} \tag{2.40}$$

where $\tilde{C}_n$ is a subset of the full set of alternatives, should give similar estimates, since under IIA, exclusion of alternatives does not affect the consistency of estimators. As it was said, it's possible to use:

- Hausman-McFadden test (Hausman & McFadden 1984).
  We have to build the following statistic:

  $$(\hat{\beta}_{\tilde{C}} - \hat{\beta}_C)'(\sum_{\hat{\beta}_{\tilde{C}}} - \sum_{\hat{\beta}_C})^{-1}(\hat{\beta}_{\tilde{C}} - \hat{\beta}_C) \tag{2.41}$$

  that is asymptotically χ^2 distributed with $\tilde{K}$ degrees of freedom, where $\tilde{K}$ is the number of elements in the subvector of coefficients that is identifiable from the restricted choice set model. So the null hypothesis that IIA holds is rejected if the value that comes from the equation 2.41 is bigger than the tabulated value of χ^2.

- McFadden, Tye and Train. In this case it's possible to build an approximate likelihood ratio test statistic with $\tilde{K}$ degrees of freedom: $-2[\ell_{\tilde{C}}(\hat{\beta}_C) - \ell_{\tilde{C}}(\hat{\beta}_{\tilde{C}})]$, where the two log likelihood values are calculated on the estimation sample for the restricted choice set model. This statistic is not a proper likelihood ratio test because $\hat{\beta}_C$ is not a vector of constants. For this reason we can consider

the following correction:

– Small and Hsiao. To remove the bias they proposed to use:

$$\frac{1}{1 - N_1/(\alpha N)}\{-2[\ell_{\tilde{C}}(\hat{\beta}_C) - \ell_{\tilde{C}}(\hat{\beta}_{\tilde{C}})]\}$$

where N is the number of observations in the unrestricted choice set estimation, N_1 is the number of observations in the restricted choice set estimation ($N_1 < N$) since those observations with chosen alternatives not in the restricted choice set are omitted, and $\alpha \geq 1$ is a scalar. Asymptotically this corrected likelihood ratio statistic actually is χ^2 distributed with $\tilde{K}$ degrees of freedom. Sometimes the assumption made by this correction that a scalar difference between the covariance matrices exists is not defensible, so it was proposed an exact test for the IIA assumption. To perform the test Small and Hsiao randomly divided the full estimation data set into two parts, denoted A and B. On sample A, using the restricted choice sets, estimated $\hat{\beta}_C^A$, the subvector of coefficients corresponding to the parameters that are identifiable when the restricted set of alternatives are used; next, on sample B, using the restricted choice set, estimated $\hat{\beta}_C^B$ and the corresponding log likelihood, $\ell_{\tilde{C}}^B(\hat{\beta}_C^B)$; finally again on sample B, but now based on the unrestricted choice sets, they obtained $\hat{\beta}_C^B$. They showed that if we form the following convex combination:

$$\hat{\beta}_C^{AB} = (1/\sqrt{2})\hat{\beta}_C^A + (1 - 1/\sqrt{2})\hat{\beta}_C^B \tag{2.42}$$

and use it to evaluate the log likelihood of the sample B with the restricted choice sets, denoted as $\ell_C^B(\hat{\beta}_C^{AB})$, then the statistic $-2[\ell_{\tilde{C}}^B(\hat{\beta}_C^{AB}) - \ell_{\tilde{C}}^B(\hat{\beta}_C^B)]$ is asymptotically χ^2 distributed with $\tilde{K}$ degrees of freedom, $\tilde{K}$ being the common dimension of the $\hat{\beta}_C^A$, $\hat{\beta}_C^B$, $\hat{\beta}_{\tilde{C}}^B$, $\hat{\beta}_C^{AB}$ parameter vectors. This test is more computa-

tionally intensive and time-consuming. Generally it's better that the simpler corrected approximate likelihood ratio test be carried out first, and then, only if its underlying assumption is violated, should the exact test procedure be used.

2.6.2 Lagrange multiplier test of IIA

This test checks if cross-alternative variables enter the model. If so, IIA assumption is violated. The test is composed of 3 steps.

- **Step 1**: We estimate the systematic utilities ($\hat{V}_{in}$) and fitted choice probabilities ($\hat{P}_n(i|C_n)$) using all N observations: $\hat{V}_{in} = \hat{\beta}' x_{in}\ \forall i \in C_n\ \hat{P}_n(i|C_n) = \frac{e^{\hat{\beta}' x_{in}}}{\sum_{j \in C_n} e^{\hat{\beta}' x_{jn}}}$
- **Step 2**: For a given $A_n \subset C_n$ we calculate auxiliary variables in the following way:

$$\hat{V}_{A_n n} = \frac{\sum_{j \in A_n} \hat{V}_{jn} \hat{P}_n(j|C_n)}{\sum_{j \in A_n} \hat{P}_n(j|C_n)} \qquad n = 1, ..., N \tag{2.43}$$

$$Z_{in}^{A_n} = \begin{cases} V_{in} - \hat{V}_{A_n n}, & if \quad i \in A_n \\ 0, & otherwise \end{cases} \quad n = 1, ..., N \tag{2.44}$$

  Since Z is non zero only for the alternatives in the set A, it contains information regarding the other alternatives in A_n. The spirit of the proposed test is to verify the presence of cross alternatives variables.

- **Step 3**: We can now estimate:

$$\hat{P}_n(i|C_n) = \frac{e^{\hat{\beta}' x_{in} + \gamma^A Z_{in}^A}}{\sum_{j \in C_n} e^{\hat{\beta}' x_{jn} + \gamma^A Z_{jn}^A}} \tag{2.45}$$

The hypothesis are the following: $H_0 : \gamma^A = 0$
$H_1 : \gamma^A \neq 0$
γ^A is distributed as a χ^2.

If we reject H_0 we reject the IIA assumption. If H_0 is not rejected nest A is considered to satisfy IIA.

2.7 Estimation of Multinomial Logit Model

The logit model has some special properties that under certain circumstances greatly simplify estimation of the parameters. Most of this theory are tributable to McFadden (McFadden 1974). The technique generally used to estimate a logit model is the Maximum Likelihood, applied as follows. We can indicate with N the sample size and define:

$$y_{in} = \begin{cases} 1, \, if \;\; n \; chose \; i \\ 0, \, otherwise \end{cases} \tag{2.46}$$

The likelihood function for a general multinomial choice model is:

$$\ell^* = \prod_{n=1}^{N} \prod_{i \in C_n} P_n(i)^{y_{in}}, \tag{2.47}$$

where for a linear in parameters logit:

$$P_n(i) = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}}. \tag{2.48}$$

Taking the logarithm of 2.47, we seek a maximum to:

$$\ell = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left(\beta' x_{in} - ln \sum_{j \in C_n} e^{\beta' x_{jn}} \right) \tag{2.49}$$

Setting the first derivatives of ℓ with respect to the coefficients equal to zero, we obtain the necessary first-order conditions:

$$\frac{\partial \ell}{\partial \hat{\beta}_k} = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left(x_{ink} - \frac{\sum_{j \in C_n} e^{\beta' x_{jn}} x_{jnk}}{\sum_{j \in C_n} e^{\beta' x_{jn}}} \right) = 0, for \quad k = 1, ..., K \tag{2.50}$$

Or in more compact form:

$$\sum_{n=1}^{N} \sum_{i \in C_n} [y_{in} - P_n(i)] x_{ink} = 0, \quad for \quad k = 1, ..., K. \tag{2.51}$$

The second derivatives are given by:

$$\frac{\partial^2}{\partial \hat{\beta}_k \partial \hat{\beta}_l} = -\sum_{n=1}^{N} \sum_{i \in C_n} P_n(i) \left[x_{ink} - \sum_{j \in C_n} x_{jnk} P_n(j) \right] \cdot \left[x_{inl} - \sum_{j \in C_n} x_{jnl} P_n(j) \right] \tag{2.52}$$

Under some weak conditions ℓ in equation 2.49 is globally concave, so if a solution to equation 2.51 exists, it's unique. The Maximum likelihood estimator of β is consistent, asymptotically normal and asymptotically efficient. The first order conditions (2.51) can be rewritten as:

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} x_{ink} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i \in C_n} P_n(i) x_{ink}, \quad k = 1, ..., K. \tag{2.53}$$

This means that the average value of an attribute for the chosen alternatives is equal to the average value predicted by the estimated choice probabilities. In particular, if an alternative-specific constant is defined

for an alternative i, then at the maximum likelihood estimates,

$$\sum_{n=1}^{N} y_{in} = \sum_{n=1}^{N} P_n(i) \tag{2.54}$$

implying that the sum of the choice probabilities for alternative i equals the number in the sample that choose i.

2.8 Nested logit

As we said in the previous sections, sometimes the IIA property doesn't hold and so we cannot apply a Multinomial Logit Model. In this circumstances we must use some other models, for instance a Nested Logit Model. In this model the set of alternatives can be partitioned in two subsets, the nests, and the following properties must hold:

- For any two alternatives in the same nest, the ratio of the probabilities is independent of the attributes of all other alternatives, so the IIA holds within the nest;
- For any two alternatives in different nests, the ratio of the probabilities can depend on the attributes of other alternatives. So in this case the IIA property doesn't hold for alternatives in different nests.

To understand the difference between Multinomial Logit and Nested Logit we can consider the figure 2.1:

We can see on the left that for a Multinomial Logit we have three alternatives on the same level, instead in Nested Logit we can have j alternatives partitioned into K nonoverlapping subsets that we can call: $B_1, B_2, ..., B_K$ and that we call nests. The utility that person n obtains from alternative j in nest B_k is denoted as: $U_{jn} = V_{jn} + \varepsilon_{jn}$.

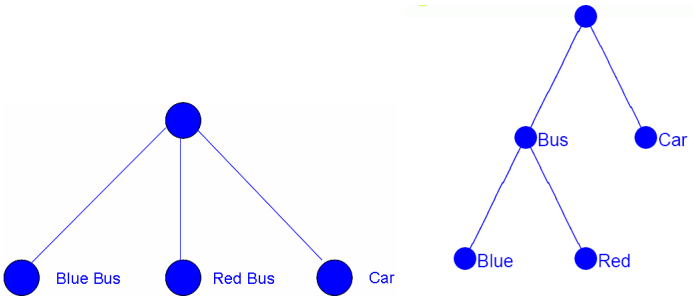

Figure 2.1: *Differences between Multinomial Logit and Nested Logit model*

The nested logit model is obtained by assuming that the the vector of unobserved utility, $\varepsilon_n = (\varepsilon_{1n}, ..., \varepsilon_{Jn})$ has cumulative distribution:

$$exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{-\varepsilon_{jn}/\lambda_k}\right)^{\lambda_k}\right) \tag{2.55}$$

This distribution is a type of GEV distribution that we will explain better in the next section. The marginal distribution of each ε_{jn} is univariate extreme value, but for any two alternatives j and m in the same nest B_k ε_{jn} is correlated with ε_{mn}. Instead for any two alternatives in different nests, the unobserved portion of utility is still uncorrelated:
$Cov(\varepsilon_{jn}, \varepsilon_{mn}) = 0 \;\; \forall j \in B_k, \;\; m \in B_l \;\; l \neq k$
The parameter λ_k is a measure of the degree of independence in unobserved utility among the alternatives in nest k and $1 - \lambda_k$ is a measure of correlation. When $\lambda_k = 1$ for all k, the nested logit model reduces to the standard logit model.
For a nested logit model the choice probability for alternative i in the nest B_k is the following:

$$P_{in} = \frac{e^{V_{in}/\lambda_k}(\sum_{j\in B_k} e^{V_{jn}/\lambda_k})^{\lambda_k - 1}}{\sum_{l=1}^{k}(\sum_{j\in B_l} e^{V_{jn}/\lambda_l})^{\lambda_l - 1}} \tag{2.56}$$

In this way is possible to show that IIA holds within each subset of alternatives but not across subsets. In fact if we consider two alternatives $i \in B_k$ and $m \in B_l$, since the denominator of 2.56 is the same for all alternatives, the ratio of the probabilities is the ratio of the numerators:

$$\frac{P_{in}}{P_{mn}} = \frac{e^{V_{in}/\lambda_k}(\sum_{j\in B_k} e^{V_{jn}/\lambda_k})^{\lambda_k - 1}}{e^{V_{mn}/\lambda_l}(\sum_{j\in B_l} e^{V_{mn}/\lambda_l})^{\lambda_l - 1}} \tag{2.57}$$

If $k = l$, so if i and m are in the same nest, then the factors in parentheses cancel out and we obtain:

$$\frac{P_{in}}{P_{mn}} = \frac{e^{V_{in}/\lambda_k}}{e^{V_{mn}/\lambda_l}} \tag{2.58}$$

It's clear that this ratio is independent of all other alternatives. If $k \neq l$, so if i and m are in different nests, the factors in parentheses do not cancel out and therefore the ratio of probabilities depends on the attributes of all alternatives in the nests that contain i and m. This ratio, however doesn't depend on the attributes of alternatives in nests other than those containing i and m. In this circumstances we can assert that a form of IIA holds and we can define it as Independence from Irrelevant Nests (IIN).

The equation 2.56 probably is not very clear, but it's useful because the choice probabilities can be expressed in an alternative way that is readily interpretable. In fact we can decompose the observed part of utility in two parts: the first one that we can indicate with W that is constant for all alternatives within a nest, and a part labeled Y that varies over alternatives within a nest. So we can rewrite the utility as follows:

$$U_{jn} = W_{kn} + Y_{jn} + \varepsilon_{jn} \quad for \quad j \in B_k \tag{2.59}$$

where:
W_{kn} depends only on variables that describe nest k. These variables differ over nests but not over alternatives within each nest.
Y_{jn} depends on variables that describe alternative j. These variables vary over alternatives within nest k.
In this way we can write the nested logit probability as the product of two standard logit probabilities:

$$P_{in} = P_{in|B_k} P_{nB_k} \tag{2.60}$$

where $P_{in|B_k}$ is the conditional probability of choosing alternative i given that an alternative in nest B_k is chosen, and P_{nB_k} is the marginal probability of choosing an alternative in nest B_k. This decomposition is useful because can be showed that the two probabilities take the form of a logit:

$$P_{nB_k} = \frac{e^{W_{kn}+\lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{ln}+\lambda_l I_{ln}}} \tag{2.61}$$

$$P_{in|B_k} = \frac{e^{Y_{in}/\lambda_k}}{\sum_{j\in B_k} e^{Y_{jn}/\lambda_k}} \tag{2.62}$$

where

$$I_{kn} = ln \sum_{j\in B_k} e^{Y_{jn}/\lambda_k} \tag{2.63}$$

The parameters of a nested logit model can be estimated by standard maximum likelihood techniques, but also in a sequential fashion, exploiting the fact that the choice probabilities can be decomposed into marginal and conditional probabilities that are logit. In this second case there are two difficulties. First, the standard errors of the upper-

model parameters are biased downward. This happens because the variance of the inclusive value estimate that enters the upper model is not incorporated into the calculation of standard errors. Second, sometimes, some parameters appear in several submodels and so, estimating the models separately provides separate estimates of whatever common parameters appear in the model. This doesn't happen with maximum likelihood, but sometimes in simultaneous estimation some problems arise and so it could be useful to estimate the model sequentially and then use the sequential estimates as starting values in a simultaneous estimation.

2.9 Cross-Nested Logit

Nested logit model is one of the solution for the problem related with the IIA property, but in many occasions we don't know if an alternative belongs only to a nest or to more than one nest. We can consider for example the choice of a transportation mode. We can divide the different mode according to the following figure: but we can also sup-

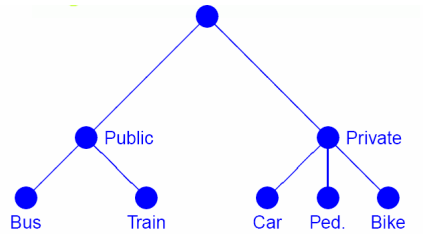

Figure 2.2: *A possible structure of Nested Logit model for transportation system*

pose a different division in nests as, for example in figure 2.4 What of the two structures is right? The solution is in the Cross-Nested logit that is an extension of the Nested Logit model (Ben-Akiva &

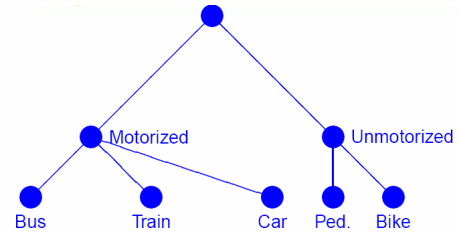

Figure 2.3: *An alternative structure of Nested Logit model for transportation system*

Bierlaire 1999),(McFadden 1978), where each alternative may belong to more than one nest. So, in this situation we can classify, for example, the transportation modes in the following way:

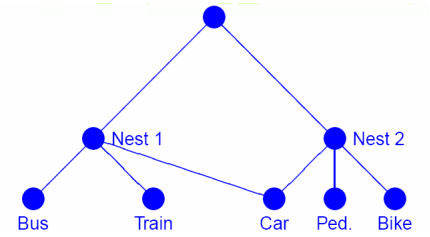

Figure 2.4: *A structure of Cross-Nested Logit model for transportation system*

Similar to the Nested Logit Model, the choice set C_n is partitioned into M nests C_{mn}. Moreover, for each alternative i and each nest m there are parameteres α_{im} $(0 \leq \alpha_{im} \leq 1)$representing the degree of membership of alternative i in nest m. The utility of alternative i is then given by:

$$U_{imn} =_{in} +\tilde{\varepsilon}_{in} +_{C_{mn}} +\tilde{\varepsilon}_{C_{mn}} + ln\alpha_{im} \tag{2.64}$$

The error terms $\tilde{\varepsilon}_{in}$ and $\tilde{\varepsilon}_{C_{mn}}$ are independent. The error terms $\tilde{\varepsilon}_{in}$ are independent and identically Gumbel distributed. The distribution

of $\tilde{\varepsilon}_{C_{mn}}$ is such that the random variable $max_{j_{mn}} U_{jmn}$ is Gumbel distributed with scale parameter μ. The probability for indivdual n to choose alternative i is given by:

$$P(i|C_n) = \sum_{m=1}^{M} P(C_{mn}|C_n) P_n(i|C_{mn}) \tag{2.65}$$

where

$$P(C_{mn}|C_n) = \frac{e^{\mu V_{Cmn}}}{\sum_{l=1}^{M} e^{\mu V_{Cln}}} \tag{2.66}$$

$$P(i|C_{mn}) = \frac{\alpha_{im} e^{\tilde{V}_{in}}}{\sum_{j_{mn}} \alpha_{jm} e^{\tilde{V}_{jn}}} \tag{2.67}$$

and

$$V_{C_{mn}} = \tilde{V}_{C_{mn}} + ln \sum_{j_{mn}} \alpha_{jm} e^{\tilde{V}_{jn}} \tag{2.68}$$

In this way we introduce in the model the level of membership of the alternatives for the different nests. The CNL model is then appealing to capture complex situations in where correlations cannot be handled by the Nested Logit. Also this model can be derived from GEV models that we will see in the next section.

2.10 GEV models

As we said previously, nested and cross-nested logit can be obtained from the GEV family (McFadden 1978), that is a general formulation from which we can derive different kind of logit model.

If we consider a function G that depends on Y_j for all j, we can

denote this function in the following way:

$$G = G(Y_1, ..., Y_j) \tag{2.69}$$

Furthermore we can indicate with G_i the derivative with respect to Y_i:

$$G_i = \frac{\partial G}{\partial Y_i} \tag{2.70}$$

If G satisfies the following conditions:

- $G \geq 0$ for all positive values of Y_j ;
- G is homogeneous of degree one;
- $G \to \infty$ as $Y_j \to \infty$ for any j;
- The cross partial derivatives of G change in the following way: $G_i \geq 0$ for all i, $G_{ij} =_i /_j \leq 0$ for all j, $G_{ijk} =_{ij} /_k \geq 0$ for any distinct i, j, k and so on for higher-order cross-partials;

then

$$P_i = \frac{Y_i G_i}{G} \tag{2.71}$$

is the choice probability for a discrete choice model that is consistent with utility maximization. Any model that can be derived in this way is a GEV model. This models are important because a purely mathematical approach allows the researcher to generate models that he might not have developed while relying only on his economic intuition. Obviously the difficulty is that the researcher has little guidance on how to specify a function G that provides a model that meets the needs of this research.

Chapter 3

Spatial issues in Discrete Choice Models

This chapter will focus the attention prevalently on the residential choice models and destination choice models, in which there are distinctive features that distinguish them from non-spatial choice problems (Pellegrini & Fotheringham 2002). Failure to account for these features may lead to erroneous analytical results and ineffective spatial policies. So it's important to consider these features that are not found typically in non-spatial models. These characteristics (Guo 2004) can be summarized as follows:

– **Definition of alternatives**: Contrary to most aspatial contexts, spatial choice problems often involve choice elements that are difficult to define (Lerman 1983). For example tourists choosing a holiday destination may be selecting among one or more different geographical levels, such as a hotel, a city, or a country. Similarly when a person chooses where to shop, we don't know if he chooses

a specific store, a neighborhood populated with shops or a specific shopping mall. Then, the definition of the choice set is far from trivial for such applications.

- **Definition of choice set**: In spatial choice situations, decision makers often face a very large set of potential options. However, in practice, the number of alternatives actually considered is constrained by the individual's limited capacity for gathering and processing information. So it seems unrealistic to assume individuals can evaluate all possible alternatives at any one time. The identification of individual choice sets is therefore a challenge to the analyst (Kanaroglou & Ferguson 1998).
- **Substitutability among choice alternatives**: Due to the continuity of space, the spatial alternatives faced by decision makers are likely to follow the First Law of Geography (Tobler 1970), that everything is related to everything else, but closer things are more closely related. An alternative at a given location may be perceived as more similar, and therefore more substitutable, to an alternative closer by rather than farther away. The perceived similarity between neighboring spatial alternatives are often intangible or difficult to quantify. Failure to account for such perceived similarity would lead to inaccurate interpretation of choice behavior. Furthermore, in standard discrete choice models, accommodating unobserved similarity among the choice alternatives is not a straightforward task.
- **Measurement of spatial variables**: As in the case of other modeling efforts, the success of a discrete choice modeling exercise relies on correct model specifications, which are tied closely to accurate representation or measurement of relevant variables. For variables that are spatial in nature, their value can be observed only after a location has been specified or a space been

demarcated. In the latter case, the continuity of space renders almost infinitely many ways for an analyst to define areal units for measuring. Without knowing which of the many spatial configurations to use, past efforts of spatial choice modeling typically use administrative spatial units, such as census tracts, for which data are readily available. These administrative units often bear no relation to how the decision makers themselves measure, or perceive, the spatial factors in their mind. Such a practice may easily lead to inaccurate analytic outcomes.

The goal of this chapter is then to present some possible techniques in which spatial problems are considered. Some of them will be showed in details in next chapter where some possible new procedures will be proposed too. For our purposes it's important to introduce before some spatial measure that will be useful for the proceeding of the chapter and of the dissertation.

3.1 Spatial weights matrix

A spatial weights matrix is defined as the formal expression of spatial dependency between observations (Anselin 1988). Research on spatial weights matrix that can be indicated with W has been reviewed in past years (Griffith 1996). Five rules of thumb can aid the specification of weights matrix:

- It's better to posit some reasonable geographic weights matrix than to assume independence. This implies that one should search for or theorize about an appropriate W and that better results are obtained when distance is taken into account.
- It's best to use surface partitioning that falls somewhere between a regular square and a regular hexagonal tessellation. For planar

data is suggested a specification between four and six neighbors.

- A relatively large number of spatial units should be employed (generally $n > 60$).

- It's always wise to choose less complicated models when the opportunity present itself.

- In general it's better to apply a somewhat under-specified rather than an over specified weights matrix, because overspecification reduces the power of tests (Florax & Rey 1995)

In the classical formulation this matrix, indicated with W, is a positive square matrix with elements w_{ij}. The cells contain the values 0 or 1 if we build a contiguity matrix, otherwise they can assume also other values. It's then important the definition of neighbors and for this reason different methods have been developed to build spatial weights matrix; let's consider some of the most important.

The Spatial Contiguous Neighbors is one of the simplest and most used method to build a contiguity matrix. Really it can be divided in three sub-techniques that we will see in details:
Rook Contiguity The four neighbors of each cell in the cardinal directions are given the value 1, all others 0. This is the most popular formulation of W. To understand better the situation is possible to consider the figure 3.1.

In the previous figure the neighbors of the cell number 5 are the number 2,4,6 and 8. According to this figure we can build the Spatial weights matrix in the following way:

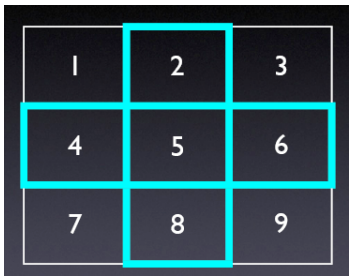

Figure 3.1: *Rook contiguity*

$$W = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

It's possible to see that the value 1 appears for the neighbor cells, otherwise the value is equal to 0.

Bishop Contiguity In this case the units sharing a vertex with a cell i are considered as neighbors of i.

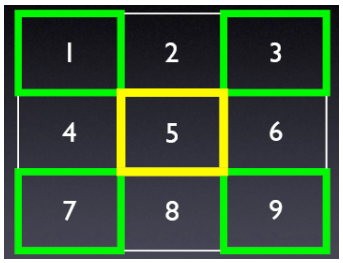

Figure 3.2: *Bishop contiguity*

Queen contiguity This last method combines the rook and bishop definition as any unit sharing a common edge or vertex with i, defined as a neighbor of i, as it's possible to see in the figure 3.3.

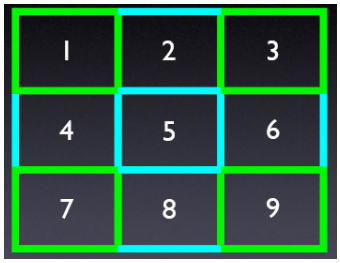

Figure 3.3: *Queen contiguity*

Obviously also for these last two techniques the building of the Spatial weights matrix follows the same procedure. Other used approaches are:

- k **nearest neighbor distance**: according to this definition all units among the k nearest neighbors of unit i are treated as neighbors of i, while the $k+1, ..., k+n$ units are treated as non-neighbors. Clearly the value k should be theoretically informed;
- **Distance band** in which the element w_{ij} of the matrix W is equal to 1 if d_{ij} is inferior to a distance cut-off; otherwise is equal to 0;
- **Cliff-Ord weights** in which $w_{ij} = [d_{ij}]^{-a}[b_{ij}]^{b}$ where d_{ij} is the distance between i and j and b_{ij} is the share of common boundary between i and j in the perimeter of i;
- **Inverse distance weights** In this case $w_{ij} = 1/d_{ij}^{\alpha}$ where d_{ij} is the Euclidean distance and α is a positive number; the most used values of α are 1,2 or 5;
- **Block structure** in which $w_{ij} = 1$ for all i and j in the same block. The blocks are defined according to some specific criterion.

Other measures that it's possible to consider are:

- Geostatistics functions (spherical, Gaussian, exponential)
- Lengths of shared borders
- Number of links

The spatial weights matrices will be useful in the proceeding of the dissertation.

A positive G_i^* indicates that there is clustering of high values around i; a negative number scrutinized cumulatively, rather than by distance bands, around each observation as absolutely with distance, the cluster diameter is reached, implying that distance. It's possible to indicate with

3.2 Spatial Multinomial Logit

Recent studies in travel behavior research focus on activity location and spatial interaction of activities. These spatial interactions and dependencies (spatial autocorrelation) warrant modeling techniques that explicitly account for space. The introduction of spatial component is important for the properties of the estimations, in fact it's demonstrated that the elimination of territorial component from a discrete choice model leads to parameter estimations that are biased (Goetzke 2003). There are two different approaches to introduce the spatial dependency: the first one consider the influence that every individual can have on the other decision makers; in the second approach the similarity between the alternatives is considered. It's then important consider the Spatial Autocorrelation, defined as the dependency found in a set of cross-sectional observations over space. It occurs when individuals in population are related through their spatial location (Anselin 1988).

To describe the first approach, in which the interactions between individuals are considered we can follow the Spatial Multinomial Logit (Mohammadian & Kanaroglou 2003). To understand how to introduce the spatial dependency

we have to consider the basic formulation of utility:

$$U_{in} = V_{in} + \varepsilon_{in} \tag{3.1}$$

As we said in the previous chapter we can divide the utility in a deterministic component V_{in} and a randomly distributed unobserved component ε_{in} capturing the uncertainty. In order to account for spatial dependency, it is assumed that the systematic component of utility function V_{in} consist of two parts; the first part is a linear in the parameter function that captures the observed attributes of decision-makers n and alternatives i, while the second term captures spatial dependencies across decision-makers. Utility of alternative i for the decision maker n is given as:

$$U_{in} = V_{in} + \varepsilon_{in} = \left(\sum \beta_i X_{in} + \sum_{s=1}^{S} \rho_{ns} y_{si} \right) + \varepsilon_{in} \tag{3.2}$$

where parameters β_i make up a vector of parameters corresponding to X_{in}, the vector of observed characteristics of alternative i and decision-maker n. Parameters ρ make up a matrix of coefficients representing the influence that the choice of decision maker s has on decision-maker n while choosing alternative i. S is the number of decision-makers who have influence on n. y_{in} will be set equal to unity if the decision-maker s has chosen alternative i, and zero otherwise. ρ can be modeled similar to an impedance function. In spatial statistics it usually takes the form of a negative exponential function of the distance separating the two decision-makers (D_{ns}).

$$\rho_{nsi} = \lambda exp\left(-\frac{D_{ns}}{\gamma} \right) \tag{3.3}$$

where λ and γ are parameters to be estimated. The total influence that the choices of all other decision-makers have on decision-maker n can be modeled

as:

$$Z_{in} = \sum_{s=1}^{S} \rho_{nsi} y_{si} \tag{3.4}$$

The probability that decision-maker n would choose alternative i rather than any other alternative j in the choice set, can be expressed as the probability that the utility of i is higher than that of any other alternative, conditional on knowing the systematic utility V_{jn} for all j alternatives in the choice set. To estimate the parameters we have to consider the following log-likelihood function:

$$L^*(\beta) = ln(L(\beta)) = \sum_{n=1}^{N} \sum_{i \in C_n} ln P_{in}^{y_{in}} = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left[V_{in} - ln \left(\sum_{j \in C_n} exp(V_{jn}) \right) \right] \tag{3.5}$$

To calculate the spatial dependency term ρ we need estimates of the parameters λ and γ. The value of these two parameters can be estimated directly by maximizing the previous maximum likelihood function or, alternatively, they can be obtained via a search procedure over a range of numbers by trying out different values of the parameter γ while estimating the value of λ as a standard parameter in logit model. The Spatial Multinomial Logit can be extended considering not only the characteristics of the decision-maker, but also variables relatives at the same time to the individual and to the chosen alternative (Nelson, Pinto, Harris & Stone 2004). We can define this model as a Conditional Spatial Multinomial Logit in which the linear component of systematic utility is decomposed in two parts as follows:

$$x_{ij}\beta_j = h_{ij}\eta + g_i\delta_j \tag{3.6}$$

in which h_{ij} includes the characteristics of every individual relative to the alternative j; instead g_i are the individual attributes that don't depend on the chosen alternative. Therefore we can write the utility in the following

way:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \left(\sum_j h_{ij}\eta + \sum_j g_i \delta_j + \sum_{k=1}^{K} \rho_{ikj} y_{kj} \right) + \varepsilon_{ij} \tag{3.7}$$

The estimation of the parameters is the same of the Spatial Multinomial Logit.

The two previous models allow us to consider the spatial effects, but they cannot solve the problem related to the IIA property. As we saw in the first chapter possible solutions are nested logit models, but if we want to introduce also the spatial component we can consider some models based on the Mixed Multinomial Logit Models.

3.3 Mixed Multinomial Logit Models

Mixed logit is a higly flexible model that can approximate any random utility model (McFadden & Train 2000). This model has been known for many years but it has only become fully applicable since the advent of simulation. A mixed logit model is any model whose choice probabilities can be expressed in the following form (Train 2003):

$$P_{in} = \int L_{in}(\beta) f(\beta) d\beta \tag{3.8}$$

where L_{in} is the logit probability evaluated at parameters β:

$$L_{in}(\beta) = \frac{e^{V_{in}\beta}}{\sum_{j=1}^{J} e^{V_{in}(\beta)}} \tag{3.9}$$

and $f(\beta)$ is a density function. If $V_{in}(\beta)$ is linear in β the probability can be expressed as:

$$P_{in} = \int \left(\frac{e^{\beta' x_{in}}}{\sum_j e^{\beta' x_{jn}}} \right) f(\beta) d\beta \tag{3.10}$$

Considering this last expression is evident that the standard logit is a special case of mixed logit, when the $f(\beta)$ distribution degenerate at fixed parameters b: $f(\beta) = 1$ if $\beta = b$ and $f(\beta) = 0$ if $\beta \neq b$. The mixed logit probability can be derived from utility-maximizing behavior in several ways. The most straightforward derivations are based on random coefficients or error components.

Random coefficients If we consider this first approach we can consider the classical expression for the utility:

$$U_{jn} = \beta_n' x_{jn} + \varepsilon_{jn}$$

This is the same specification as for the standard logit except that β varies over decision makers rather than being fixed. So it's necessary to specify a distribution for the coefficients and estimates the parameters of that distribution. In most applications (Revelt & Train 1998), (Ben-Akiva & Bolduc 1996) $f(\beta)$ has been specified to be normal or lognormal: $\beta(b, W)$ or $ln\beta(b, W)$ with parameters b and W that are estimated. However also other distributions have been used as, for example, triangular and uniform distributions (Revelt & Train 1998).

Error components In this second approach the model represent error components that create correlations among the utilities for different alternatives. Utility is then specified as:

$$U_{jn} = \alpha' x_{jn} + \mu_n' z_{jn} + \varepsilon_{jn} \tag{3.11}$$

Here the terms in z_{jn} are error components that, along with ε_{jn}, define the stochastic portion of the utility. It can be indicated as $\eta_{jn} = \mu_n' z_{jn} + \varepsilon_{jn}$ and it can be correlated over alternatives depending on the specification of z_{jn}. When z_{jn} is identically 0, then the model reduces to a standard logit model. In this approach the emphasis is placed on specifying variables that can induce correlations over alternatives. Before considering the introduction of the spatial component in the model it's important to underline that any Random Utility Model (RUM) can be approximated to any degree of accuracy by a mixed logit with appropriate choice of variables and mixing distribution (McFadden & Train 2000).

3.4 Discrete choice models and spatial dependence

Introduction of Mixed logit model has been very important because it allowed to consider the spatial contiguity between destination and not only between individual. Really one of the first approach in which the spatial weights have been introduced (Autant-Bernard 2005) starts from a simple Multinomial Logit Model in which the contiguity is introduced in the deterministic part of the utility:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \rho_1 W_1 V_{ij} + X\beta + \varepsilon \tag{3.12}$$

where the matrix W_1 is one of the spatial weights matrix introduced previously and ρ_1 are the parameters relative to the spatial effects. Successively, however, according to a different approach, the introduction of the spatial component is not in the deterministic part, but in the random part of a Mixed Multinomial Logit Model:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = X\beta + \eta + \xi = X\beta + \eta + \rho_2 W_2 \xi + \upsilon \tag{3.13}$$

Here the random component is divide in two parts:

η with mean equal to 0 and Gumbel distributed;
ξ with mean equal to 0 and distribution equal to $f(\xi|\Omega)$ where Ω is the matrix of the parameters of the distribution, depending on the structure of the observed data and on the alternatives of the choice set.
ξ here is an autoregressive component with υ vector of normally distribute variables with mean = 0 and variance-covariance matrix = to Σ. If we define $\Sigma = \sigma^2 I$, it's possible to write $\upsilon = \sigma\tau$ where σ indicate the standard deviation and τ is a vector of elements normally standardized distributed. Following the two showed techniques it is possible to consider an integrated approach (Miyamoto, Vichiensan, Shimomura & Paez 2004) in which the spatial component was introduced to a double level: in the deterministic part and in the random part. In this way the special effect in choice probability is accommodated considering the spatial interaction among the observable data and the spatial autocorrelation among the unobservable data:

$$U = V + \varepsilon \begin{cases} V = \rho_1 W_1 V + X\beta \\ \varepsilon = \eta + \xi = \eta + \rho_2 W_2 \xi + \sigma\tau \end{cases} \tag{3.14}$$

In this way the utility can be re-expressed as follows:

$$U = (I - \rho_1 W_1)^{-1} + \sigma(I - \rho_2 W_2)^{-1}\tau + \varepsilon \tag{3.15}$$

In this equation $(I-\rho_1 W_1)^{-1}$ represents the spatial interaction among observable data,$\sigma(I - \rho_2 W_2)^{-1}\tau$ is the Spatial autocorrelation among unobservable data, while ε is IID Gumbel distributed. The choice probability can then be expressed as:

$$L(\tau) = \frac{exp\left((I - \rho_1 W_1)^{-1} + \sigma(I - \rho_2 W_2)^{-1}\tau\right)}{\sum exp\left((I - \rho_1 W_1)^{-1} + \sigma(I - \rho_2 W_2)^{-1}\tau\right)} \tag{3.16}$$

The parameters to be estimated in the model include scalars ρ_1 and ρ_2 representing the degree of spatial dependency, the standard deviation σ and the vector β associated with the explanatory variables in the deterministic part of the model. We can define a parameter vector θ that includes all parameters in the model. Estimation can be done with the classical maximum likelihood method, which has commanded substantial attention in recent years (Bhat & Guo 2004). In particular if we write the log-likelihood function:

$$L(\theta) = \sum_n \sum_i y_{ni} log L(\theta) \tag{3.17}$$

and considering the 3.18:

$$y_{in} = \begin{cases} 1, if\ n\ chose\ i \\ 0, otherwise \end{cases}$$

we obtain that the log likelihood in 3.17 involves the evaluation of multidimensional integrals that are not in closed form. Simulation techniques anyway are useful to approximate the multidimensional integrals and maximizing a simulated log-likelihood function (Bhat 1998). The simulation techniques entail computing the integral at several values of τ drawn from a the normal distribution for a given value of the parameter vector θ and averaging the integrand values. The choice probabilities are approximated by averaging over the NR numbers of simulated probability (SP):

$$SP = \frac{1}{NR} \sum_{nr=1}^{NR} \frac{exp\left((I - \rho_1 W_1)^{-1} + \sigma (I - \rho_2 W_2)^{-1} \tau\right)}{\sum exp\left((I - \rho_1 W_1)^{-1} + \sigma (I - \rho_2 W_2)^{-1} \tau\right)} \tag{3.18}$$

The above expression is an unbiased estimator of the actual probability. The simulated log-likelihood (SLL) function is the following:

$$SLL = \sum_n y ln(SP) \tag{3.19}$$

An ulterior possibility to introduce the spatial correlation is when we consider a different model (Ben-Akiva, Bolduc & Walker 2001) named Logit Kernel. Also in this case the random part can be divided in two components: a probit-like term with a multivariate distribution and a Gumbel random variate. The probit-like term captures the interdependencies among the alternatives. These interdependencies can be specified using a factor analytic structure (McFadden 1984):

$$\varepsilon_n = F_n \xi_n + \upsilon_n \tag{3.20}$$

where ξ_n is an $(M * 1)$ vector of M multivariate distributed latent factors, F_N is a $(J_n * M)$ matrix of the factor including fixed and unknown parameters and may also be a function of covariates, and υ_n is a $(J_n * 1)$ vector of Gumbel random variates. For estimation it's desirable to specify the factors such that they are independent and so ξ_n can be decomposed as follows:

$$\xi_n = T\zeta_n \tag{3.21}$$

where ζ_n are a set of standard independent factors, TT' is the covariance matrix of ξ_n and T is the Cholesky factorization of it. Here if we consider a generalized autoregressive process of the errors ξ, we have $\xi = \rho_2 W_2 \xi + T\zeta$ and the utility is expressed as:

$$U = X\beta + \eta + \rho_2 W_2 \xi + T\zeta \tag{3.22}$$

that is similar to the 3.13

3.5 Aggregation of alternatives

In the previous sections we have seen how to introduce the spatial component in logit models, but the proposed techniques don't solve the problems related

to the size of choice set. In fact when the number of alternatives in the choice set is large, response probability models may impose heavy burdens of data collection and computation. In this section we introduce one of the possible solution for this issue, *aggregation of alternatives*.

The use of the grouped alternatives model to approximate the ideal disaggregate models was, at beginning, forced on the analysts simply because data were not available for all the alternatives at the original level of the elemental alternatives (Lerman 1983). Yet, although over the years more micro-level data have become available, residential choice studies of a disaggregate nature remain scarce. This is perhaps because few researchers have risen to challenge the norm, i. e., the aggregate approach, but also because the concept of grouped alternatives has its behavioral merits. According to this approach, we can consider an individual who select an alternative and we can indicate with C the set of all possible alternatives; the aggregation amounts to partitioning this set into C_i subsets that do not overlap (Parsons & Needelman 1992):

$$C_i \subseteq C, \quad i = 1, ..., J$$

where each i is an aggregate alternative. The choice probability of an aggregate alternative is equal to the probability that the individual chooses one of its elemental alternatives. So we can write it as follows:

$$P_n(i) = \sum_{l \in C_i} P_n(l), \quad i = 1, ..., J \tag{3.23}$$

Now we can consider again the individual's utility for an elemental alternative:

$$U_{ln} = V_{ln} + \varepsilon_{ln} \tag{3.24}$$

where V_{ln} is the classical deterministic part of utility. The utility of choosing an aggregate alternative i is just:

$$U_{in} = max(V_{ln} + \varepsilon_{ln} | l \in C_i) \tag{3.25}$$

so U_{in} is the maximum utility that the individual n perceive among all alternatives in the group i. If the ε_{ln} are independent and identically distributed Gumbel random variables with location parameter equal to 0 and scale parameter μ, we can decompose 3.25 as:

$$U_{in} = \frac{1}{\mu} ln \left[\sum_{l \in C_i} e^{\mu V_{ln}} \right] + \varepsilon_{in} \tag{3.26}$$

where $\frac{1}{\mu} ln \left(\sum_{l \in C_i} e^{\mu V_{ln}} \right)$ is the mode of the random variable $max(V_{ln} + \varepsilon_{ln}), l \in C_i$ and ε_{in} is Gumbel distributed with mode 0 and scale scale μ. We can now decompose 3.26 as follows:

$$\begin{aligned}
U_{in} &= \frac{1}{\mu} ln \left[\sum_{l \in C_i} e^{\mu V_{ln}} \right] + \varepsilon_{in} \\
&= \frac{1}{\mu} ln \left[\sum_{l \in C_i} e^{\mu \bar{V}_{in}} e^{\mu (V_{ln} - \bar{V}_{in})} \right] + \varepsilon_{in} \\
&= \frac{1}{\mu} \left[ln(e^{\mu \bar{V}_{in}}) + ln \sum_{l \in C_i} e^{\mu (V_{ln} - \bar{V}_{in})} \right] + \varepsilon_{in} \\
&= \bar{V}_{in} + \frac{1}{\mu} \left[ln \sum_{l \in C_i} e^{\mu (V_{ln} - \bar{V}_{in})} \right] + \frac{1}{\mu} ln M_i + \frac{1}{\mu} ln \left(\frac{1}{M_i} \right) + \varepsilon_{in} \\
&= \bar{V}_{in} + \frac{1}{\mu} \left[ln \left[\frac{1}{M_i} \sum_{l \in C_i} e^{\mu (V_{ln} - \bar{V}_{in})} \right] \right] + \frac{1}{\mu} ln M_i + \varepsilon_{in} \\
&= \bar{V}_{in} + \frac{1}{\mu} ln B_i + \frac{1}{\mu} ln M_i + \varepsilon_{in}
\end{aligned} \tag{3.27}$$

where $\bar{V}_{in}$ is the average utility of the elemental alternatives in aggregate alternative i; $B_i = \frac{1}{M_i} \sum_{l_i} e^{\mu(V_{ln} - \bar{V}_{in})}$ is a measure of the heterogeneity of the elemental choices and M_i is the number of disaggregate alternatives in the aggregate one i. Sometimes, when there is not a sufficient information, site aggregation schemes usually specify the utility of an aggregate alternative ignoring the terms involving lnB_i and lnM_i and $\bar{V}_{in}$ or some approximate measure is used alone and there is a loss in estimation accuracy due to these omissions. A drawback of aggregation is the fact that we need a procedure to join together the different alternatives. Different techniques have been proposed in past years.

The most common practice to aggregate the alternatives in residential choice problems or destination choice, is to join dwellings or destinations into administratively defined units, typically census tracts or transport analysis zones. The tracts or zones are then considered as the communities or neighborhoods that the individual households choose from. Other administratively defined units used as proxy for residential alternatives include counties (Gabriel & Rosenthal 1989), school districts and census cities (Levine 1998). The use of administrative units is likely attributed to the fact that spatial data describing the residential environment of the dwellings are often readily available only for these units. Anyway there are also other kind of geographical aggregations; for example analysts divide the study area into 0,5 by 0,5 mile-squares-zones. Data about the housing quality and neighborhood are then aggregated over these "quarter-sections" (Anas & Chu 1984).

One of the problem that arises when aggregation of zones is effected is the Modifiable Area Unit Problem (MAUP). The effect of the MAUP has been found in a variety of spatial analysis and modeling studies, including univariate statistical analyses, bivariate regression, multivariate statistical analysis. While relevant research effort has concentrated mostly on revealing the MAUP, the search for effective solutions has not been widely attempted, at least not with satisfactory results. We can categorize the past attempts in

three categories (Wong 1996):

- Data Manipulation;
- Technique oriented;
- Error modeling.

The data manipulation approach is based on the suspicion that the MAUP would vanish if the chosen areal units can be justified one way or another, instead for administrative convenience. Many researchers developed methods for creating optimal zones with respect to predefined objective functions.
The technique-oriented approach, on the other hand, is based on the argument that the MAUP effect might have been a result of using inappropriate models or statistical techniques in analyzing aggregated spatial data. This leads to the proposal of abandoning the unsuitable classical statistical techniques and replacing them with frame independent analysis (Tobler 1991).
Another group of researcher recognize that, when analysis moves from one spatial scale to another, relationships among variables and among spatial entities also change. Instead of searching for techniques immune to such scale effects, they adopt the error modeling approach of explicitly documenting variations derived from changing scale, and incorporating these changes into modeling and analysis.
Generally we can assert that to reduce or remove the effect of MAUP it's necessary to know something about the general nature of the phenomenon. In temporal instances there are often strong organizing principles associated with the observations that give rise to self-similarity, which analyst can exploit to perform generalization. What often makes the spatial instances difficult is the lack of intuition about the phenomenon at hand and analysts are thus required to decide on the spatial units before attempting to study the phenomenon.
Instead of using locality-based groupings, some studies construct residential choice alternatives by grouping individual units based on their non-

spatial attributes; sometimes, i.e., different housing types have been defined as choice alternatives (Quigley 1976). Other times the neighborhood types, defined based on tract-level median income values, are considered as choice alternatives in examining individual's preferences for neighborhood qualities (Chattopadhyay 2000) and similarly, communities defined based on census places are grouped into clusters based on residential density and commute time to form location choice for individuals.
In next chapter we will introduce a technique for carrying out the aggregation that will consider spatial and non-spatial characteristics simultaneously.

3.6 Problems related to the size of the choice set

Aggregation of alternatives is only one method to reduce the number of the alternatives, but sometimes, after the aggregation there is the application of a nested logit model and so, really, the number of alternatives is not reduced and then the computation is anyway not so easy. To handle this problem, it's possible to follow another procedure. It was demonstrated (McFadden 1978), in fact, that if the multinomial logit functional form is valid, consistent estimates of the parameters of the strict utility function can be obtained from a fixed or random sample of alternatives from the full choice set. We can denote with C the full choice set and with $P(i|C, x, \beta)$ the true selection probabilities where β is a vector of parameters and x a vector of explanatory variables. If the IIA property is satisfied we can write the choice probabilities as follows:

$$i \in D \subseteq C \Longrightarrow P(i|C, x, \beta) = P(i|D, x, \beta) \sum_{j \in D} P(j|C, x, \beta) \qquad (3.28)$$

where D is a subset drawn from the set C according to a probability distribution $\pi(D|i,x)$, which may, but need not, be conditioned on the observed choice i. The observed choice may be either in or out of the set D. We can consider different examples of π distributions:

- choose a fixed subset D of C, independent of observed choice
- choose a random subset D of C, independent of observed choice
- choose a subset D of C, consisting of the observed choice i and other alternatives selected randomly

The most used type is the last one and we can show some examples of this:

To obtain consistent estimator from a sample of alternatives we must introduce the *positive conditioning property* established by McFadden (McFadden 1978):
If $j \in D \subseteq C$ and $\pi(D|i,x) > 0$ then $\pi(D|j,x) > 0$

and the *uniform conditioning property*:
If $i, j \in D \subseteq C$, then $\pi(D|i,x) = \pi(D|j,x)$.

If the two previous properties are respected, then the maximization of the modified likelihood function:

$$L_n = \frac{1}{N} \sum_{n=1}^{N} log \left[\frac{e^{V_{i_n}(x_n,\beta)+log\pi(D_n|i_n,x_n)}}{\sum_{j \in C} e^{V_{j_n}(x_n,\beta)+log\pi(D_n|j_n,x_n)}} \right] \tag{3.29}$$

yields, under normal regularity conditions, consistent estimates of the unknown parameters.

As showed sampling of techniques is an applied technique for reducing the computational burden involved in estimating a choice model with a large number of alternatives. Now the main issue is how to obtain the most effective sample of alternatives. Many strategies are possible like:

- Simple Random Sampling of Alternatives;
- Importance Sampling of Alternatives;
- Independent Importance Sampling;
- Importance Sampling with Replacement;
- Stratified Importance Sampling;

We will discuss about some of these.

3.6.1 Simple Random Sampling of Alternatives

The simplest and most used approach to sample design is to draw a simple random sample of alternatives from the full choice set and to add the chosen alternative if it is not otherwise included. The probability to obtain a subset D is

$$\pi_n(D|i) = \left(\begin{array}{c} J \\ J' \end{array} \right)^{-1}, \quad i \in D \tag{3.30}$$

where the term in parenthesis indicates the number of combinations of J items taken J' at a time. A problem that could rise with this kind of sampling is the fact that if the observed choice is sampled, the size of the sampled choice set, that we can indicate with $\tilde{J}_n$, is equal to J'; if the observed choice is not sampled, $\tilde{J}_n$ is equal to $J'+1$. To prevent this drawback it's possible to draw randomly J' alternatives from all the available alternatives, except for the chosen one. In this case the set D has always $J'+1$ elements and:

$$\pi_n(D|i) = \left(\begin{array}{c} J-1 \\ J' \end{array} \right)^{-1}, \quad i \in D \tag{3.31}$$

These random sampling strategies are characterized by the *uniform conditioning* and so the correction terms for alternative sampling bias in the logit

model $ln \hookrightarrow \pi_n(D|i), i \in D$ are equal and therefore cancel out in the choice probabilities and a standard logit model with a choice set given by D yields consistent estimates.

3.6.2 Importance Sampling of Alternatives

With a simple random sampling the alternatives may have very small choice probabilities and so a sample of alternatives in which the alternatives most likely to be chosen by the decision maker have a higher probability of being selected may be more efficient. This is the basic idea of importance sampling that is borrowed from Monte Carlo integration. We can, for example, consider the probability of estimating a sum of choice probabilities over a subset with J_0 alternatives:

$$\sum_{i=1}^{J_0} P_n(i)$$

It's efficient to select a sample from the J_0 alternatives with selection probabilities q_{in} such that the ratios

$$\frac{P_n(i)}{q_{in}}, \qquad i = 1, ..., J_0$$

vary as little as possible. Thus an importance alternative sampling strategy is based on preliminary estimates of the choice probabilities. These estimates can be provided a priori by some simple model form. For example, in destination choice model two factors are usually considered, distance and size, which may be combined in a gravity-type function:

$$\tilde{M}_i e^{-\alpha d_{in}}, \quad i = 1, ..., J \tag{3.32}$$

where $\tilde{M}_i$ is an approximate measure of size of destination zone i, d_{in} is a measure of distance between the origin of traveler n and destinations in zone i, and α is a scalar parameter that represents the sensitivity to distance. The drawback of this kind of sampling is that it's only an intuitively reasonable strategy for model estimation.

3.6.3 Stratified Importance Sampling

The technique of stratified importance sampling avoids the need to specify a selection probability q_{in} for every alternative $j = 1, ..., J$. The set of J alternatives is stratified into R disjoint subsets such that:

$$\sum_{r=1}^{R} J_{rn} = J \tag{3.33}$$

where J_{rn} is the number of alternatives in stratum r for decision maker n. The importance sampling criterion is realized by assigning different selection probabilities in different strata, while maintaining uniform selection probabilities in different strata. So if we indicate with $\tilde{J}_{rn}$ the sample size for stratum $r = 1, ..., R$ and with $r(i)$ the stratum of alternative i we must draw a simple random sample of size $\tilde{J}_{rn}$ from every stratum except that from the stratum of the chosen alternative i, from which we draw only a sample of $\tilde{J}_{r(i)n} - 1$ alternatives and then we add the chosen alternative. In this way the size of D is uniform across all observations. The probability of selecting a set of alternatives D is:

$$\pi_n(D|i) = \begin{pmatrix} J_{r(i)n} - 1 \\ \tilde{J}_{r(i)n} - 1 \end{pmatrix}^{-1} \prod_{r=1} \begin{pmatrix} J_{rn} \\ \tilde{J}_{rn} \end{pmatrix}^{-1}, \qquad i \in D \tag{3.34}$$

The main advantages of this method are:

- its fixed sample size ($\tilde{J} = J' + 1$ where J' is the total number of random draws
- the selection probabilities, given by $q_{in} = \frac{\tilde{J}_{r(i)n}}{j_{r(i)n}}$ are easier to quantify than in the other methods

In next chapter we will show how it's possible the use of Multidimensional Analysis to build a new kind of Stratified Sampling. Multidimensional Analysis will be also the basis of the new technique of aggregation that we pointed out previously.

Chapter 4

Multidimensional Analysis and Residential Choice Models

In this chapter we will introduce, as we said, two methods to deal with the great number of alternatives in the choice set and with the spatial complexity related to residential choice. The basis of the two methods will be Principal Component Analysis that we will introduce briefly at beginning of the chapter.

4.1 Principal Component Analysis

The central idea of Principal Component Analysis is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in data set (Jolliffe 2002). This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. So we have to consider our original data and the various step to obtain the Principal Components. First of all we have to define with X the

data matrix in which there are the values of different variables for all the individual of the analysis; with D we indicate the matrix of the weights for every unity and with M the metric matrix, that defines the nature of the distances between individual. The matrix $X_{n,p}$ define a group of vectors that describe a cloud of points and the distances between them define the shape of the cloud. To see this shape we can project it on some spaces, having has objective the minimization of the deformation caused by the projection. So we can indicate with u a unit vector of the p-dimensional space R^p and with OH_i the orthogonal projection of the individual OM_i on the line generated by the vector u. The situation can be understood looking at the following figure 4.1.

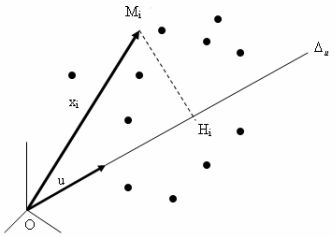

Figure 4.1: *Projection of individuals on an optimal subspace*

OH_i can be expressed as follows:

$$OH_i = x_i' M u \tag{4.1}$$

and the objective is therefore to search, according to the Ordinary Least Squares, the line that minimizes the sum of the square distances of the points that we can indicate with $\sum_i^n (M_i H_i)^2$. But it's easy to see that:

$$\sum_{i=1}^{n} (M_i H_i)^2 = \sum_{i=1}^{n} (OM_i)^2 - \sum_{i=1}^{n} (OH_i)^2 \tag{4.2}$$

and so, as $\sum_{i=1}^{n} (OM_i)^2$ is independent from the vector u, to minimize the

quantity $\sum_{i=1}^{n}(M_iH_i)^2$ is the same to maximize $\sum_{i=1}^{n}(OH_i)^2$, that can be expressed as follows:

$$max_{(u)}\left\{\sum_i p_i OH_i^2\right\} = max_{(u)}\{u'MX'DXMu\} = max_{(u)}\{u'Au\} \quad (4.3)$$

with the normalization constraint equal to $u'Mu = 1$; p_i are the weights of the individuals and $A = MX'DXM$. We can then consider the Lagrange Multiplier and write:

$$L = u'Au - \lambda(u'Mu - 1) = max \quad (4.4)$$

If we derive respect to u we obtain:

$$\frac{\partial L}{\partial u} = 2Au - 2\lambda Mu = 0 \quad (4.5)$$

and then:

$$Au = \lambda Mu \quad (4.6)$$

Following some algebraic manipulations we can obtain:

$$\begin{aligned} u'Au &= \lambda u'Mu \\ \lambda &= u'Au \end{aligned} \quad (4.7)$$

and if M is a positive definite matrix, it's possible to write:

$$\begin{aligned} M^{-1}Au &= \lambda u \\ A^* &= M^{-1}A \\ A^*u &= \lambda u \end{aligned} \quad (4.8)$$

So finding the eigenvector u_1 associated with the first eigenvalue of the ma-

trix M^{-1} we can calculate the first principal component: $c_1 = XMu_1$. To obtain the other components is only necessary to introduce a constraint of orthogonality $u_1'Mu_2$=0 and so we have:

$$\frac{\partial L}{\partial u_2} = 2Au_2 - 2\lambda_2 Mu_2 - \iota_2 Mu_1 = 0 \tag{4.9}$$

If we multiply all the elements for u_1' we have $u_1'Au_2 = u_1'Mu_2 = 0$ and $u_1'Mu_1 = 1$ and so we obtain $Au_2 = \lambda_2 Mu_2$ and therefore the second component is derivable from the eigenvector associated to the second eigenvalue of the same matrix as before. The importance of Principal Component Analysis is due to the fact that, as we said at beginning of the chapter, it allows to consider only few components to explain most of the variability present in the data. In the continuing of the chapter we will introduce some modification to the classical Principal Component Analysis and we will show how the techniques based on it can be useful for our purposes.

4.2 Constrained Principal Component Analysis

In classical Principal Component Analysis there is the hypothesis of a symmetrical relationship between variables, but sometimes, this hypothesis can be not assumed and so it's necessary consider some different techniques. The problem of the asymmetrical relationship was faced. i. e., through a sort of "visualized regression" by means of the supplementary points technique (Lebart, Morineau & Warwick 1984), but one of the most important proposed techniques is the Constrained Principal Component Analysis (D'Ambra & Lauro 1982). This technique allows to use the Principal Component Analysis also when the relationships between the variables involved are not symmetrical. In Constrained Principal Component Analysis we indicate with

X and Z the data matrices associated respectively with two sets of quantitative variables $x_j(j = 1, ..., p)$ and $z_k(k = 1, ..., q)$ observed on the same individuals:

$$Z = \begin{bmatrix} z_{11} & ... & z_{iq} \\ ... & ... & ... \\ z_{n1} & ... & z_{nq} \end{bmatrix} X = \begin{bmatrix} x_{11} & ... & x_{iq} \\ ... & ... & ... \\ x_{n1} & ... & x_{nq} \end{bmatrix}$$

In this technique we don't assume that the variables play a symmetrical role in the analysis, but instead a non-symmetrical one. The aim of CPCA is to analyze the structure of explained variance of the first data set due to the second one, assumed as explanatory. We can indicate with R_x and R_z the two sub-spaces of R^n spanned by the linearly independent vectors x_j and x_z. Constrained Principal Component Analysis consists in carrying out a Principal Component Analysis of the image of x_j obtained onto R_z through a suitable orthogonal projection operator. We can obtain it in the following way:

$$P_z = Z(Z'D_pZ)^{-1}Z'D_p \tag{4.10}$$

where D_p is the diagonal metric matrix with $1/n$ on the diagonal. We use this operator because he gives us the best image of X matrix on R_z according to the ordinary least squares. Now we can project X on R_z:

$$X^* = P_zX \tag{4.11}$$

and we can effectuate the principal component analysis in the following way:

- Research of the subspace of reduced dimensions (principal axis) of R_z through the computation of eigenvalues and eigenvectors of the following expression:

$$X'P_zXv_\alpha = \lambda_\alpha v_\alpha \tag{4.12}$$

where $\lambda_\alpha > 0$, $v_\alpha \; v'_\alpha = 0$, $\alpha \neq \alpha'$

– Determination of principal factors as follow:

$$u_\alpha = \frac{1}{\sqrt{\lambda}_\alpha} v_\alpha \tag{4.13}$$

with $u'_\alpha \; u_\alpha$=1

– Research of the principal components as linear combination of the original variables x:

$$c_\alpha = P_z X v_\alpha \tag{4.14}$$

The norm of these components is equal to λ_α and so, to obtain normalized components $c'_\alpha c_\alpha$ it's necessary to divide the previous eqaution for $1/\sqrt{\lambda_\alpha}$:

$$c_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \tag{4.15}$$

– Computation of the correlation between original variables and new component to describe the relationships between the two groups of variables and to interpret the components. This computation is possible if we multiply the 4.15 for X':

$$X' c_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X' P_z X v_\alpha \tag{4.16}$$

and considering 4.12 we have

$$X' c_\alpha = \sqrt{\lambda_\alpha} \tag{4.17}$$

It's also possible to calculate the correlations of the new components with the variables of the group Z:

$$Z'c_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}(Z'X)v_\alpha = Z'Xu_\alpha \tag{4.18}$$

This expression shows that CPCA allows to analyze the image of the external correlation between the two sets of variables X and Z on the principal axes, unlike the supplementary points technique where x_j variables are independently projected on the factorial axes, without participating to the determination of them. Also in this technique anyway it's possible to represent some supplementary variables in the following way:

$$y_s'P_zc_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}y_s'P_zXv_\alpha \tag{4.19}$$

To find a supplementary individual we can consider instead the following expression:

$$c_s = \frac{1}{\sqrt{\lambda_\alpha}}z_s(Z'Z)^{-1}z_s'x_sv_\alpha \tag{4.20}$$

Our idea to deal with the great number of alternatives will be based on Constrained Principal Component Analysis, but before showing it, we present some other techniques used in the past years to consider an integration between spatial and multivariate analysis, introducing also the links with classical univariate indexes.

4.3 Univariate indexes of spatial structure

We saw in the previous chapter that it's possible to indicate with W a contiguity matrix and we indicate, as usually, with X the data matrix and furthermore we can write $q_i = x_i - \bar{x}$ with $\bar{x} = \frac{1}{n}\sum_i^n x_i$, where n is the

number of elements of our analysis. This notation will be useful to introduce in this paragraph Geary index and Moran index that are the basis of spatial statistics.

4.3.1 Moran Index and Geary contiguity coefficient

Moran index can be defined in the following way:

$$I = \frac{n \sum w_{ij} q_i q_j}{\sum w_{ij} \sum_{i=1}^{n} q_i^2} \tag{4.21}$$

where n is the number of statistic elements. Sometimes this matrix is written as follows:

$$I = \frac{q' F q}{\sum_{i=1}^{n} q_i^2 / n} \tag{4.22}$$

where the elements of the matrix F are $f_{ij} = \frac{w_{ij}}{\sum_{ij} w_{ij}}$. Generally Moran index is used in three different fields, we are interested in the case of the neighboring graph. In this circumstance the index can be written as follows:

$$I = \frac{1}{2w} \frac{q' W q}{\sum_{i=1}^{n} q_i^2 / n} \tag{4.23}$$

where W is the contiguity matrix and $2w$ is the number of the pairs of neighbors. We have that $I_n' W I_n = 2W$ and so the previous expression can be written as 4.22. The c coefficient of Geary instead, is generally known as:

$$c = \frac{\sum w_{ij} (x_i - x_j)^2}{2_{ij} \sum_{i=1}^{n} q_i^2 / (n-1)} \tag{4.24}$$

or sometimes it's written as

$$c = \frac{f_{ij}(x_i - x_j)^2}{2\sum_{i=1}^{n} q_i^2/(n-1)} \quad or \quad c = \frac{\frac{1}{2w}{}_{ij}(x_i - x_j)^2}{2\sum_{i=1}^{n} q_i^2/(n-1)} \tag{4.25}$$

In this index we use the term $1/(n-1)$ for the variance rather than $1/n$ that is used for Moran index. The substantial difference between the two indexes is perhaps that Moran index is arranged so that its extremes match the intuitive notions of positive and negative correlation, whereas the Geary index uses a more confusing scale.

4.4 Statistical analysis of contiguity

The first attempt to introduce spatial component in multivariate analysis has been done by Lebart (Lebart 1969). To understand the analysis we have to introduce a neighboring graph as in the following figure:

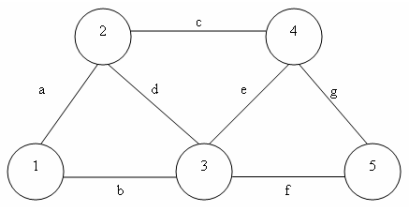

Figure 4.2: *graph of neighboring*

Here we have some edges that join the individual. We can then introduce the following matrices:

- $W = [w_{ij}]$ is, as usually, the symmetric n by n matrix of the between-sites neighbors: if alternative i is neighboring alternative j then $w_{ij} = 1$, else $w_{ij} = 0$. Moreover for any i, $w_{ii} = 0$
- N is the diagonal matrix of degrees of vertices with $n_i = \sum_j m_{ij}$

In our example we can rewrite the matrix as follows:

$$\text{W}= \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \mathbf{N} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

The matrix $N - W$ can be indicated as a proximity operator. If now we indicate with $w = \sum_i n_i = \sum_{ij} m_{ij}$ twice the number of edges we can consider a matrix T with $w/2$ rows and n columns, crossing the $w/2$ edges and the n elements. If an edge k joins two vertices i and j and if $i < j$, $t_{ki} = 1$ and $t_{kj} = -1$; $t_{ki} = 0$ otherwise. For our graph the matrix of edges will be:

$$\text{T}= \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

The following relation between the matrices is straightforward:

$$T'T = N - W \tag{4.26}$$

The matrix $T'T$ is a symmetric and semi-definite positive matrix because $(x'L'Lx \geq 0)$, then also $N - W$ is semi-definite positive. It's important to remember this property because it will be useful for our new proposal. However, before showing it, we can demonstrate the relationships with the Geary and Moran indexes; to do this, we must introduce:

- $P = [p_{ij}]$ with $p_{ij} = \frac{1}{2w} w_{ij}$ where w is the total number of pairs of neighbors, therefore $\sum_{ij} p_{ij} = 1$
- $D = Diag(p_1, p_2, ..., p_n)$ is the diagonal matrix of neighboring weights: $p_i = \frac{1}{2w} \sum_j w_{ij}$

We have now all the matrices we need to define the total variance, the local variance and the global variability (Thioulouse, Chessel & Champely 1995). In fact if we consider the mean of a variable x, given the weights D, is equal to:

$$\bar{x}_D = \sum_i p_i x_i = x' D I_n \tag{4.27}$$

It's variance is equal to what we can call *total variance*:

$$Var(x) = \sum_i p_i (x_i - \bar{x}_D)^2 \tag{4.28}$$

If x is D-centered it can be written in matrix form as:

$$Var(x) = x' D x \tag{4.29}$$

The *local variance* (Aluja & Lebart 1984) is:

$$LV(x) = \sum_i \sum_j p_{ij} (x_i - x_j)^2 \tag{4.30}$$

and can be written as:

$$LV(x) = x'(D - P)x = x' D (I_n - D^{-1} P) x \tag{4.31}$$

The *global variability* or *spatial auto − covariance* is defined by:

$$GV(x) = \sum_i \sum_j p_{ij} (x_i - \bar{x}_D)(x_j - \bar{x}_D) \tag{4.32}$$

which, if x is D-centered, can be written:

$$GV(x) = x'Px = x'D(D^{-1}P)x \tag{4.33}$$

Since it is not always positive, it cannot be called global variance. The second form in 4.31 and 4.33 show that the global variability can be seen as the covariance between x and the mean of its neighbors, and that the local variance can be seen as the covariance between x and the difference between each point and the mean of its neighbors. It's possible to derive then a variance decomposition of the following form:

$$Var(x) = LV(x) + GV(x) \tag{4.34}$$

When the neighboring weights D are uniform, the ratio of the local variance to the total variance $LV(x)/Var(x)$ is equal, except a $(n-1)/n$ factor to Geary's coefficient of autocorrelation from which Geary's index can be deduced. Similarly it's possible to note that Moran's index is exactly, under the same hypothesis, the ratio of the global variability to the total variance: $GV(x)/Var(x)$.

Lebart introduced the spatial component in Multivariate Analysis, generalizing the concept of local variance and obtaining a Local Principal Component Analysis, simply with the diagonalization of the spatial covariance matrix $X'(D-P)X$. Le Foll (LeFoll 1982) used a similar approach; he asserted in fact that the local structure of data matrix can be accomplished by the analysis of the triplet $(X_D, I_p, D-P)$ where X_D is the D-centered matrix. The row scores of this analysis maximize the local variance. Wartenberg instead (Wartenberg 1985) presented a method called Multivariate Spatial Correlation Analysis (MSCA), based on the eigenvector analysis of matrix $X'WX$ (the spatial covariance matrix). By introducing the D-centering and using $P = \frac{1}{2w}W$, we simply obtain the analysis of the global structure of the data table by the PCA of the triplet (X_D, I_P, P). The row scores of this

analysis have the highest possible global variability, but the corresponding eigenvalues are not always positive, as they are not variances but spatial auto-covariances. This method is the only one we know in which the multivariate analysis is not constrained to give positive eigenvalues.
In this paragraph we have seen some techniques to integrate spatial and multivariate analysis. The matrix used by Lebart to calculate the spatial auto-covariance will be useful to demonstrate that the which one we want to introduce in CPCA is semi-definite positive too. We will show this in the following section.

4.5 A new approach to aggregate alternatives in Discrete Choice Models

In the previous chapter we said as different techniques for aggregation of alternatives have been proposed in the past years. These techniques carry out the aggregation according to some spatial measurements or according to some characteristics of the elemental alternatives. We don't know techniques in which spatial and non-spatial elements are considered together. The only attempt we know is a Spatial Zoning Algorithm (Hammadou, Thomas, Tindemans, Witlox, Hofstraeten & Verhetsel 2004) in which they use a classical Principal Component Analysis with a Varimax rotation, then a Cluster Analysis based on Ward's method to group sectors that look alike in terms of scores of the components and, eventually, they define the aggregate alternatives, keeping in mind the observed reality and grouping together the neighboring sectors. In this way they introduce the spatial dimension only in the last step of the algorithm. Our proposal instead has as goal to introduce directly in the first step of the analysis the spatial component. This is possible considering the Constrained principal Component Analysis.
In fact we can consider the expression (4.12) and multiply it for $P_z X$; in this

way we obtain:

$$P_z X X' P_z X v_\alpha = \lambda_\alpha P_z X v_\alpha \tag{4.35}$$

The projection operator is idempotent and so if we multiply P_z for P_z we obtain once again P_z. Thanks to this property it's possible to rewrite the previous equation in the following way:

$$P_z X X' P_z P_z X v_\alpha = \lambda_\alpha P_z X v_\alpha \tag{4.36}$$

In this way we can then compute the Principal Component as the eigenvectors associated to this last matrix (the principal axis of the other subspace). The matrix that we must diagonalize is therefore the following:

$$Z(Z' D_p Z)^{-1} Z' D_p X X' D_p Z (Z' D_p Z)^{-1} Z' \tag{4.37}$$

This matrix allows euclidean representation of individuals on the space on which CPCA in R^n rests. In this perspective some contiguity matrix could be considered instead that XX' (D'Ambra & Lauro 1992). In this way we introduce the spatial component directly in the multivariate analysis. One problem could be the fact that the matrix must be semi-definite positive and we cannot be sure that using a simple contiguity matrix, as those proposed in chapter 3, it will be semi-definite positive. For this reason we propose to carry out some modifications on the contiguity matrix keeping in mind the concepts of total variance, local variance and global variability that we introduced before (Lebart 1969)and that have been developed during the years (Monestiez 1978), (LeFoll 1982), (Mom 1998). The matrix that we will introduce in lieu of XX' is $F - P^*$ (Meot, Chessel & Sabatier 1993), (Cornillon, Amenta & Sabatier 1999) that we will show to be semi-definite positive too. We consider once again the contiguity matrix that we indicated with W; we can then define the matrix P^* simply multiplying the contiguity matrix for

the spatial weights matrix D_p. Now we can compute the column marginal of this new matrix; they will be the diagonal elements of the diagonal matrix F. The elements of this matrix can be indicated with $f_{ij} = \sum_j w_{ij} d_j$ where d_j are the column marginal. We can demonstrate that this matrix is semi-definite positive; in fact if we consider the proximity operator introduced by Lebart $(N - W)$ it's easy to see that, when the different zones have the same weight, the matrix $(N - W)$ is simply n *times* the matrix $F - P^*$, where n is the number of elements of the matrix. To show that this relation holds we can consider the following steps:

- we saw that P^* is obtained as $D_p W$, in this way all the elements of the new matrix will be simply the same as in W but divided for n, because the elements on the diagonal of D_p are all equal to $1/n$.
- for the same reason the matrix F will have elements equal to those of W but divided for n;
- then the relation between $N - W$ and $F - P^*$ is straightforward: $\frac{1}{n}(N - W) = (F - P^*)$ and if $(N - W)$ is semi-definite positive, as demonstrated before, also $F - P^*$ will be semi-definite positive.

When in the spatial weights matrix the elements have a different importance, and so on the diagonal there are values different from $1/n$ it's possible anyway to show that the matrix will be semi-definite positive (Cornillon, Sabatier & Chessel 1993) We can then introduce the matrix $F - P^*$ in lieu of XX' (D'Ambra, Rodia & Pagliara 2005), and so we obtain:

$$Z(Z' D_p Z)^{-1} Z' D_p F - P^{*\prime} D_p Z (Z' D_p Z)^{-1} Z' \tag{4.38}$$

Diagonalizing this matrix we obtain directly, as we showed before, the value of the components obtained calculated on the sub-space of R^n that we indicated with R_z. Once we obtained the new components, we can effectuate a cluster

analysis on them, with one of the classical methods known and, in this way we will obtain some clusters that we can define homogeneous according to multivariate analysis with a spatial constraint (D'Ambra & Lucadamo 2006). The clusters obtained can be used as aggregate alternatives in a logit model; in this way we consider, for the computation of the aggregate zones, not only the geographic characteristics but at the same time also other important variables. In the other works we know about this problem, the two approaches are always considered separately.
We will show in next chapter how this technique can be applied to a real data-set, but before, in next section, we will show that most of the properties that hold for PCA are valid also for CPCA. Afterwards we will see how the CPCA and the cluster analysis can be used also to carry out a new kind of stratified sampling of alternatives, that, as we said previously, is another method to simplify the analysis when we have a great number of alternatives in the choice set.

4.6 Properties of CPCA

Optimality of the solution

One interesting property of principal components is to give the best description of the image of X on the subspace R_z:

$$||P_z XX' - \sum c_\alpha c'_\alpha||^2 = min \tag{4.39}$$

Considering the first principal component and the trace of the matrix we can write

$$||P_z XX'||^2 + ||c_1 c'_1||^2 - 2tr(P_z XX' c_1 c'_1) \tag{4.40}$$

and so to find the minimum of the first expression it's sufficient to find

the maximum of the last term of 4.40 considering the constraint for the normalization $c_1'c_1$ we obtain the equation for the principal components as linear combinations of z_j:

$$P_z XX'zb_\alpha = \lambda_\alpha \tag{4.41}$$

Now it's possible to define a measurement for the quality of the representation

$$t_h = 1 - \frac{\sum \lambda_\alpha}{tr(X'X)} = \frac{\sum \lambda_\alpha}{tr(X'X)} \tag{4.42}$$

but if the number of z variables is inferior to the x variables this measurement cannot ever be equal to 1 so it's necessary a correction; in fact we know that the maximum of the numerator is $tr(Z'XX'Z)$ and so t_h can be rewritten as follows:

$$t_h^c = \frac{\sum \lambda_\alpha}{tr(Z'XX'Z)} \tag{4.43}$$

If we multiply the 4.41 for b_α' and we consider the normalization condition we have:

$$b'_{XX'Zb_\alpha} \tag{4.44}$$

that can be expressed also in the following way:

$$(x_1'Zb_\alpha)^2 + ... + (x_i'Zb_\alpha)^2 = \lambda_\alpha \tag{4.45}$$

In this way it's easy to see that if the variables are standardized λ_α is the sum of the correlation squares between the original variables and the principal components. The explanatory power of the principal components respect to

the original variables is given by:

$$t_h^c = \frac{{}_i'Zb_\alpha}{p} \tag{4.46}$$

where p is the number of x variables. The contribution of every variable to the components will be given by:

$$\frac{(x_i'Zb_\alpha)^2}{\lambda_\alpha} \tag{4.47}$$

Orthogonality of CPCA

To demonstrate that the Constrained Principal Components are orthogonal we can consider two eigenvalues λ_α and λ_α'. Furthermore we can write $P_zXX' = A$ and then we have:

$$\begin{aligned} AZb_\alpha &= \lambda_\alpha \\ AZb_\alpha' &= \lambda_{\alpha'} Zb_{\alpha'} \end{aligned} \tag{4.48}$$

We can do some algebraic manipulation and so we obtain:

$$\begin{aligned} b_\alpha' Z'AZb_\alpha &= \lambda_\alpha b_\alpha' Z'Zb_\alpha \\ b_\alpha' Z'AZb_{\alpha'} &= \lambda_\alpha b_\alpha' Z'Zb_{\alpha'} \end{aligned} \tag{4.49}$$

and as $Z'AZ$ is a symmetric matrix we have:

$$(\lambda_\alpha - \lambda_{\alpha'})(b_\alpha' Z'Zb_{\alpha'}) = 0 \tag{4.50}$$

and so, as we supposed that $\lambda_\alpha \neq \lambda_{\alpha'}$, the components must be necessarily orthogonal.

Other properties

In the classical analysis if we consider a transformation, not orthogonal, on the original matrix the PCA is not invariant. In the CPCA we must consider the transformations on the two groups of variables. In fact if we write $X^* = XT$ where T is a non-singular matrix and considering the classical equation we have:

$$Z(Z'Z)^{-1}Z'X^*X^{*'}g_\alpha = \psi_\alpha Z g_\alpha \tag{4.51}$$

and considering the transformation of the matrix:

$$Z(Z'Z)^{-1}Z'XTT'X'Zg_\alpha = \psi_\alpha Z g_\alpha \tag{4.52}$$

that, if TT' is not equal to the 4.36 and so it has different eigenvalues and eigenvectors. If we, instead, carry out the transformation on the matrix Z, we have $Z^* = ZS$ where S is always not singular and:

$$ZS(S'Z'ZS)^{-1}S'XX'ZSf_\alpha = \lambda ZSf_\alpha \tag{4.53}$$

and following a property of the inverse of a product of matrix we have:

$$Z(Z'Z)^{-1}Z'XX'ZSf_\alpha = \lambda_\alpha \tag{4.54}$$

and

$$P_zXX'(ZSf_\alpha) = \lambda_\alpha(ZSf_\alpha) \tag{4.55}$$

that is similar to the 4.37. The eigenvalues are in fact the same and the eigenvectors are have the following relations:

$$c_\alpha = (ZSf_\alpha) \tag{4.56}$$

and so the CPCA is invariant for a transformation on the matrix of the

variables Z.

CPCA in relation to the sub-space orthogonal and complement The Constrained Principal Analysis we showed considered the projection on the sub-space R_z, but sometimes we would like to analyze a structure of dependence of a group of variables x without the influence of the others and so we need to analyze X in the sub-space orthogonal and complement that we can indicate with $R_z^{\perp}$. The projection operator in this space is $(I - P_z)$ and so if we write:

$$\tilde{X} = (I - P_z)X \tag{4.57}$$

and remembering that $(I - P_z)^2) = (I - P_z)$, we have:

$$\begin{aligned} \tilde{X}'\tilde{X} &= X'(I - P_z)X \\ X'(I - P_z)Xd_\alpha &= \varrho_\alpha d_\alpha \end{aligned} \tag{4.58}$$

and the equation for the principal components is:

$$(I - P_z)XX'(I - P_z)Xd_\alpha = \varrho(I - P_z)Xd_\alpha \tag{4.59}$$

4.7 Cluster Sampling of alternatives

The CPCA can be used also to explain how it's possible to carry out a stratified sampling of alternatives. As we said McFadden showed that when the number of alternatives in the choice set is too large, it's possible to estimate the parameters of the model with a subset of alternatives. Different techniques were introduced in the previous chapter and generally the most used is the simple random sampling of alternatives. Sometimes the importance sampling of alternatives is used, but in these circumstances we need some particular variables as i.e. distance and size to build the so called gravity

type function and, anyway, it's only an intuitively reasonable strategy for model estimation.
Our idea, following a precedent work (Bierlaire & Lucadamo 2006), is to build the strata of our sampling according to this procedure. First of all we must carry out a CPCA on the matrix Z in which we have the values of the variables observed on the elemental alternatives. Also in this case, we introduce the spatial component instead to divide the variables in two groups. Once we obtain the components we can conduce a Cluster Analysis, taking the advantage to have components, as we showed before, that are uncorrelated. At this point we have obtained clusters of different sizes and therefore, for the sampling, we must assign a different selection probabilities in different strata, while maintaining uniform selection probabilities within strata. To define the probability for every strata we suggest two different procedures.

Probability computed according to the size of the clusters. Once we obtained the clusters, following the procedure we explained before, a simple possibility is to define the size of the sub-set we want and then proceed to the sampling in the following way:

- Let k be the number of clusters we obtain from CPCA and Cluster Analysis;
- Let define with J the number of alternatives in the full choice set;
- Let R_i be the number of alternatives in every cluster, where $i = 1, ..., k$
- Let J'_i be the size of the sub-set we defined, $i = 1, ..., k$;
- Let define with R'_i the number of alternatives we have to draw from every cluster, where $i = 1, ..., k$

then the following equality must hold:

$$\frac{R'_i}{J'_i} = \frac{R_i}{J} \tag{4.60}$$

and then:

$$R'_i = \frac{R_i}{J_i} J'_i \tag{4.61}$$

In this way we obtain a number of alternatives from every cluster that is proportional to the size of it.

Probability computed according to the size and the variability of the clusters. In the second approach we don't consider only the size of the clusters, but also the variability and so we build the probability for an alternative to be chosen following the approach proposed by Neyman.
We must indicate with σ_i the standard deviation in every cluster and then we can calculate for each of them the following measure:

$$CS_i = \sigma_i R_i \tag{4.62}$$

We can sum them for all the clusters and we obtain: $CS = \sum_i^k CS_i$. Considering the same notations as before we can write the following equality:

$$\frac{R'_i}{J'_i} = \frac{CS_i}{CS} \tag{4.63}$$

It's easy to see that if all the stratum standard deviations are equal, the optimal sample sizes are proportional to the stratum sizes and if all the stratum sizes are equal, the optimal sample sizes are proportional to the stratum standard deviations.
In next chapter we will apply the proposed techniques to a data set relative to the choice of apartment in the Zurich area. It will be showed how the proposed approaches could be useful in real situations and the results will be

compared with the traditional techniques used until now.

Chapter 5

A study in the Zurich Area

5.1 Description of the data

The data used in this analysis have been furnished by the ETH of Zurich and are fruit of the period of study of the candidate in the department ROSO - Ecole Polytechnique Federale de Lausanne. They come from two different data sources:

- Revealed preference information about households in the Greater Zurich Area that was gathered by means of an household survey conducted in 2005;
- Real estate offers that were obtained from the Web.

About the first kind of data the survey was shipped to 9330 households in 21 municipalities of Canton Zurich and surrounding cantons plus four city districts of Zurich. It contained question concerning sociodemographic features of the households, characteristics of their dwellings, and housing price information (Burgle 2006). The return rate of the survey was of 30%. These households records were geocoded, but considering that only those households that had occupied their present dwelling for no more than five years,

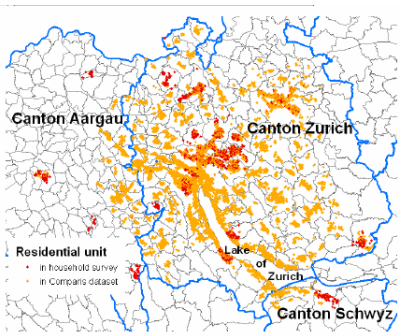

Figure 5.1: *Spatial scale of the residential choice location models*

when answering the survey, were considered recent movers and therefore eligible for the modeling of residential location choice. To complement the information collected in the household survey and to gather a reasonable amount of data to build meaningful models, real estate offers were obtained from the online real estate portal "comparis". The webpages were parsed using a Java programme. Data posted on the Internet in the period from December 2003 until October 2005 was scanned to collect a comprehensive database of real estate bids for the area in which the household survey had been conducted. For the two data sets not all the available records could be used for estimation depending on the quality of geocoding and on the variables considered in the models, because in many observations there were missing data. We will see in next paragraph the variables of the data set that we considered to build our model.

5.2 Explanatory variables

When we deal with a problem concerning the choice among a group of alternatives, the classical instrument is the Multinomial Logit Model, but we saw that there are many other models that we can apply. This is only a

first step of the analysis, in fact the main problem regard the choice of the right variables to build the model. There is a range of publications available indicating what types of variables to use for the estimation of residential location choice. Generally discrete choice models are based on the assumption that the probability for a decision maker to choose a given alternative is a function of his socioeconomic characteristics and the relative utility of alternative. The attractiveness of a residence in turn can be ascribed to attributes of the dwelling itself and attributes of its location. In our work we considered some past studies, working hypotheses on the same data and data availability.

Access to other type of opportunities, or land-use is one of the attributes that has been empirically shown to influence residential choice behavior. For example it's showed that as the amount of commercial activities increases in a zone, the probability of that zone being chosen increases. The propensity for easy access to shopping opportunities, the access to workplaces and access to alternative modes of transportation can also have an effect on residential choice.

Residential density is, sans doubt, one of the attribute that most influence a choice, but the effects are not always the same. Some works showed that households generally have an aversion to location with high density (Ben-Akiva & Bowman 1998), other researches found that high population density is preferred by households. The contradictory findings may be attributed to the difference in the population segment or the geographical area being studied. It could be also a result of other sources of error such aggregation bias.

Housing affordability, measured by housing price, or by price-income ratio is generally found to be an attractive feature for a residential zone.

Other factors that can influence a choice are **race and ethnicity**, **socioeconomic status**, **age and family status**, **school quality** and **safety**.
In our work the selection of variables was made keeping in mind not only the previous general rules, but also some precedent works on the same data set The starting working hypotheses are therefore the following:

- Households prefer to spend as little as possible of their income on housing;
- Households with employed persons prefer housing locations close to their place of employment;
- Households with children prefer to live in areas with many children;
- Young households without children prefer locations with high population density;
- Municipality characteristics like tax index or rate of vacant housing units influence residential location choice;
- Good accessibility by public transport is important for households without a car.

Keeping in mind this previous information we built our model. The full choice set was composed of 696 alternatives, because, as we said before, for many of them there were missing data. First of all we considered only the variables used in other models and we checked their explanatory power estimating a multinomial logit model. This was only a first step, because we had more than 50 variables and so there was a stepwise introduction of additional variables to test if there was an added explanatory power. At the end we had 7 significant variables for which we resume in 5.1 the description and the average.

The estimation of a Multinomial Logit Model was carried out with BIOGEME (Bierlaire 2003), (Bierlaire 2005) and the results we obtained are in

Variable	*Description*	*Mean*	*Unit*
access	Public transport accessibility for households without cars	0,90	
childdensity	Average number of children per hectare measured in a radius of 500m	1,63	person/ha
distwork	Distance between residential location and place of employment	13,96	Km
popyoung	Average number of inhabitants per hectare measured in radius of 1km multiplied for a dummy variable relative to the presence of young people	14,82	person/ha
rentratio	Total monthly rent divided for the income	0,37	
taxindex	Ratio of tax rate to the cantonal average weighted with total tax payers multiplied by total tax burden	92,06	
timetoplatz	Car travel time to Zurich centre based on regional transport model	29,63	Minutes

Table 5.1: *Description of variables considered for residential choice estimation*

the table 5.2

In the construction of the model we found that the accessibility by private or public transport showed no significant influence on residential location choice, but introducing an interaction term, representing the accessibility to population by public transport and the absence of cars in the household yielded a significant result with positive sign. These findings confirmed the assumption that accessibility only has an impact on residential location choice in connection with the availability of mobility tools in the decision-making household. The second parameter we can see in the table is the density of children per hectare for households with children under 12 years old. It has a negative sign demonstrating that households with children prefer to settle in areas where other families don't live. The distance to place of employment shows a negative sign. This results confirms the hypothesis that households prefer residential location close to the place of employment.

Variable	**Beta**	**Std Error**	**T-test**	**Rob. Std Error**	**Rob. t-test**
Access	0,5176	0,0843	6,1434	0,0818	6,3314
Childdensity	-0,0519	0,0241	-2,1572	0,0238	-2,1831
Distwork	-0,1424	0,0071	-20,1580	0,0085	-16,8473
Popyoung	0,0179	0,0020	8,8296	0,0017	10,3193
Rentratio	-1,2266	0,2603	-4,7117	0,2526	-4,8551
Taxindex	-0,015	0,0038	-3,9002	0,0037	-4,0873
Timetoplatz	0,0732	0,0062	11,7616	0,0063	11,5785

Table 5.2: *Model parameters for residential location choice*

The population density was also found to be significant if multiplied for a dummy variable relative to the presence of young people in the households. Obviously the rent ratio and the tax index have a negative influence on the utility, because they indicate a greater financial burden for the decision making household. The last parameter that we found to be significant was the travel time to Zurich centre. It has a positive sign showing that people prefer to live far from place where traffic and noise could be very high.
The model we estimated on the full choice set (all the alternatives we could use) is a basis to effect the comparison for the results we obtained with the use of the new proposed approaches.

5.3 Application of Constrained Principal Component Analysis

As we said in the previous chapter we can apply a Principal Component Analysis or a Constrained Principal Component Analysis, before carrying out a Cluster Analysis to obtain strata for a Stratified Sampling. It's easy to obtain the results for the Principal Component Analysis in the classical way, so we don't show here the procedure and we will consider only the

results. We can instead consider the steps utilized to obtain the Constrained Principal Component Analysis with the introduction of Spatial Component. It was possible to apply this technique because we had many variables that we didn't use to build the Multinomial Logit Model and that we could use for our purpose. As we showed before the first thing that we need for a Constrained Principal Component Analysis is the computation of the projector operator (4.10):

$$P_z = Z(Z'D_pZ)^{-1}Z'D_p$$

The matrix Z has on the rows all the alternatives of our analysis and on the columns all the continuous variables (56). In this case in fact, as we saw, we don't need to divide the variables in two groups, because the matrix XX' in the equation 4.37 will be substituted by the contiguity matrix. Having no more information about the alternatives, we decided to give an equal weight to all the possible choices so we have that D_p is a diagonal matrix with the value $1/n$ on the diagonal. The last matrix we needed was the proximity operator that we could calculate considering the geographic code that we had for all the apartments. We so built the contiguity matrix with a code in S-plus and then we did all the necessary steps to obtain the proximity operator. Once we introduced it in the matrix 4.37 we could carry out the CPCA and we obtained the eigenvalues that are useful to know how many components it's better to consider for the following steps of the analysis and the eigenvectors of the space R^n (the components we need for the analysis). In tables 5.3 and 5.4 we can see the differences between the eigenvalues of the classic Principal Component Analysis and of the Constrained Principal Component Analysis.

The values in the tables cannot be compared, but anyway they are useful to define the number of components to choose before applying the Cluster Analysis. We decided to consider 10 components in the two cases; in this way

Number	**Eigenvalue**	**Percentage**	**Cumulative percentage**
1	23,3871	41,76	41,76
2	4,7905	8,55	50,32
3	3,6561	6,53	56,85
4	2,9793	5,32	62,17
5	2,8777	5,14	67,30
6	2,0012	3,57	70,88
7	1,8317	3,27	74,15
8	1,7880	3,19	77,34
9	1,4847	2,65	79,99
10	1,2738	1,64	82,27
11	1,1696	2,27	84,36
12	1,0384	2,09	86,21
13	0,9192	1,85	87,85
14	0,8628	1,54	89,39
15	0,7258	1,30	90,69

Table 5.3: *Eigenvalue for the PCA*

Number	**Eigenvalue**	**Percentage**	**Cumulative percent**
1	0,0011499	35,47	35,47
2	0,0006969	21,49	56,97
3	0,0005135	15,83	72,81
4	0,0004555	14,05	86,86
5	0,0002078	6,41	93,27
6	0,0000794	2,44	95,72
7	0,0000654	2,01	97,74
8	0,0000272	0,83	98,58
9	0,0000197	0,60	99,18
10	0,0000101	0,31	99,49
11	0,0000083	0,25	99,75
12	0,0000042	0,12	99,88
13	0,0000015	0,04	99,93
14	0,0000010	0,03	99,96
15	0,0000008	0,02	99,98

Table 5.4: *Eigenvalue for the CPCA*

we had the 82% of variability explained for the classical Principal Component Analysis and the 99% for the Constrained Principal Component Analysis. With the PCA we reduced the redundant information in the data, and this advantage can be utilized in the following Cluster Analysis too.

5.4 Cluster Sampling according PCA and CPCA

The application of the Cluster Analysis is then the first step to obtain the strata from which we can extract the alternatives. Both for the PCA and for the CPCA the Cluster Analysis was effected considering the Ward's method. The number of clusters we obtained was equal to 5 in the two cases and we can see on the following figures, obtained with SPAD, how they are disposed on the first two factors.

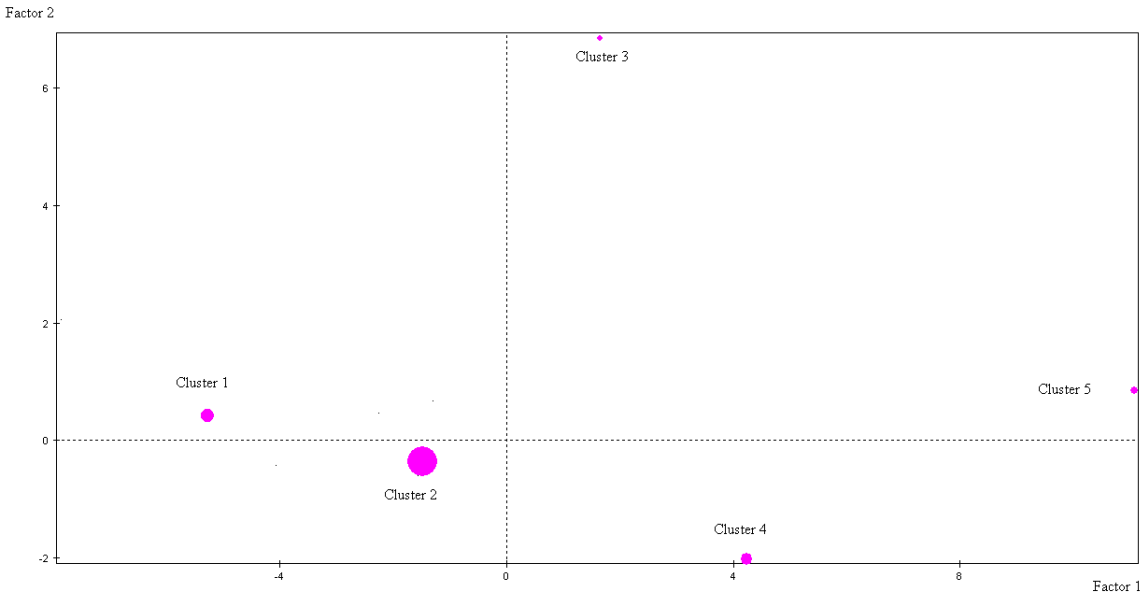

Figure 5.2: *PCA clusters*

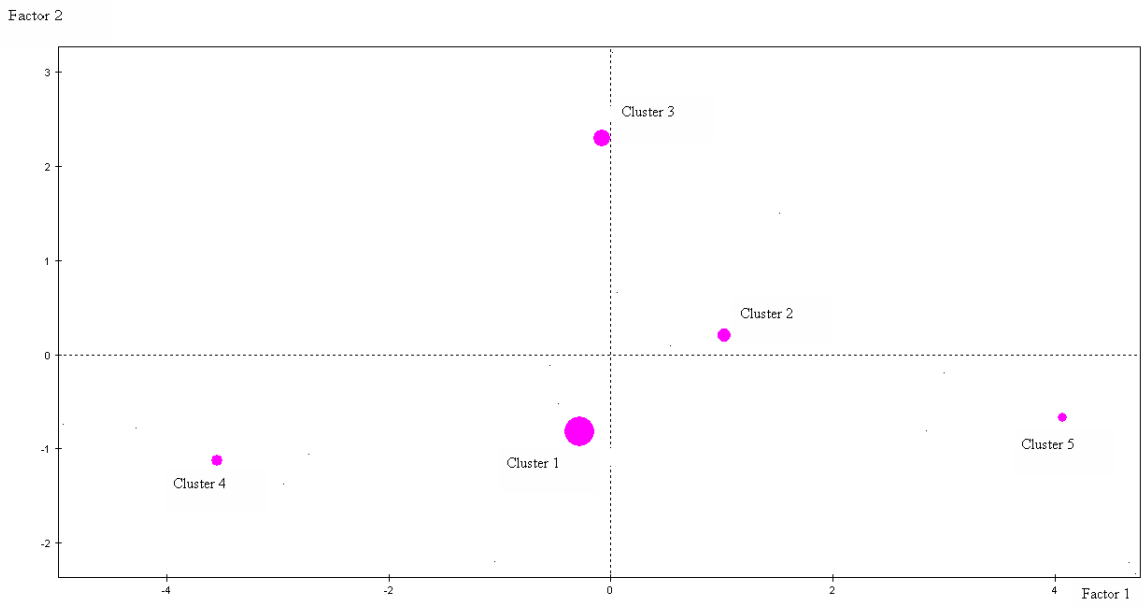

Figure 5.3: *CPCA clusters*

The differences are obviously in the composition of the clusters. As we said in the previous chapter the sampling of alternatives from clusters is done proportionally to the number of elements in every cluster. To verify if the Cluster Sampling can show better results than the Simple Random Sampling we decided to consider 5 different sample sizes for the number of alternatives: 10, 12, 15, 20, 40. Furthermore, for each sample size, the sampling procedure was repeated 5 times using different random seeds to estimate the variance due to the sampling of alternatives. In the tables 5.5 and 5.6 we indicate the number of elements we have in every cluster and the corresponding number we extract from them for every size. Therefore we obtained 5 sub-sets for every size and for every technique. On each of them we estimated the same model that we had estimated on the full choice set. We will do the same also for the classical Simple Random Sampling and, in next section, we will introduce some synthetic measures to compare the results obtained with the

Cluster number	Number of elements in every cluster	Elements drawn for every sample size				
		10	**12**	**15**	**20**	**40**
1	138	2	2	3	4	8
2	349	4	6	8	10	20
3	34	1	1	1	1	2
4	108	2	2	2	3	6
5	67	1	1	1	2	4

Table 5.5: *Number of elements to be drawn for the Cluster Sampling based on PCA*

Cluster number	Number of elements in every cluster	Elements drawn for every sample size				
		10	**12**	**15**	**20**	**40**
1	281	4	5	6	8	16
2	109	2	2	2	3	6
3	152	2	3	3	4	9
4	85	1	1	2	3	5
5	69	1	1	2	2	4

Table 5.6: *Number of elements to be drawn for the Cluster Sampling based on CPCA*

different methodologies.

5.5 Simple Random Sampling and Comparisons of the Results

The Simple Random Sampling was carried out with another code in S-plus and also in this case it was repeated 5 times to consider the variability in the results due to the sampling. Once we obtained all the results, the problem is about the comparison of them. For our purpose we followed a previous work (Nerella & Bhat 2004). The measures we considered for the evaluation of the differences between the techniques are the following:

- Ability to recover model parameters;
- Ability to replicate the choice probability of the chosen alternative for each observations;
- Ability to estimate the overall log-likelihood function accurately

For each of the criteria identified above, the evaluation of proximity was based on three properties:

- Bias, or the difference between the mean of estimates for each sample size of alternatives across the 5 runs and the true values;
- Simulation variance, or the variance in the relevant parameters across the 5 runs for each sample size of alternatives;
- Total error, or the difference between the estimated and the true values across all 5 runs for each sample size of alternatives.

Before computing all the mentioned performance measures we can have some preliminary information from the data, simply considering the significance of the parameters estimated on the different sub-sets. In figures 5.4, 5.5 and 5.6 we can see the differences in the three cases. The values we show in this part of the dissertation are relative to the samples with 20 elements, but we obtained similar results also with for other sizes.

In these tables we can already see that the probability to draw sub-sets that give estimation of the parameter not close to the true values is higher for the random sampling rather than for Cluster Sampling based on PCA or on CPCA. In this particular case, two subsets of the five extracted give a low value of the robust t-test. This is only what we see at beginning, but the following tables will show other advantages of the proposed techniques. The figure 5.7 shows the differences between the mean, across the 5 runs, of the parameters and the values estimated on the full choice set.

	Full choice set		Random 1		Random 2		Random 3		Random 4		Random 5	
Parameters	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test
b_access	0,518	6,331	0,763	5,842	0,793	5,978	0,712	5,322	0,308	4,399	0,823	5,78
b_childdensity	-0,052	-2,183	-0,042	-2,484	-0,039	-2,251	-0,025	-1,824	-0,045	-2,850	-0,013	-0,94
b_distwork	-0,142	-16,847	-0,081	-12,560	-0,084	-11,950	-0,058	-11,205	-0,095	-12,773	-0,056	-11,02
b_popyoung	0,018	10,319	0,015	9,401	0,015	9,637	0,012	7,885	0,017	11,367	0,008	4,41
b_rentratio	-1,227	-4,855	-0,894	-4,307	-0,873	-4,008	-0,673	-3,316	-0,925	-4,547	-0,932	-3,14
b_taxindex	-0,015	-4,087	-0,016	-5,667	-0,016	-5,779	-0,016	-6,477	-0,013	-4,473	-0,015	-5,85
b_timetoplatz	0,073	11,579	0,055	10,621	0,060	10,588	0,047	9,040	0,055	9,931	0,046	8,66

Figure 5.4: *Parameters estimated with the random sampling (size=20)*

	Full choice set		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
Parameters	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test
b_access	0,518	0,084	0,213	3,375	0,307	4,381	0,609	6,160	0,267	4,216	0,714	6,316
b_childdensity	-0,052	0,024	-0,044	-3,101	-0,039	-2,383	-0,047	-2,785	-0,041	-2,916	-0,045	-2,746
b_distwork	-0,142	0,007	-0,078	-13,611	-0,099	-12,431	-0,092	-11,686	-0,073	-13,102	-0,091	-11,540
b_popyoung	0,018	0,002	0,014	9,080	0,018	10,448	0,013	8,411	0,013	8,656	0,015	9,121
b_rentratio	-1,227	0,260	-0,993	-4,260	-0,811	-3,949	-0,928	-4,120	-0,984	-4,591	-0,932	-4,324
b_taxindex	-0,015	0,004	-0,010	-3,172	-0,011	-2,804	-0,012	-3,280	-0,010	-3,137	-0,015	-4,191
b_timetoplatz	0,073	0,006	0,039	7,629	0,049	7,169	0,054	8,475	0,036	7,016	0,057	8,985

Figure 5.5: *Parameter estimation with the cluster sampling on PCA (size=20)*

We can see that with the Cluster and the "Cluster CPCA" the sum of the differences between the parameters is reduced, so we have in this way a lower bias. As we said before, this happens also for the other sizes. If we look at the figure 5.8 we see that also the variance is reduced for the two new techniques.

	Full choice set		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
Parameters	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test	Value	Rob t-test
b_access	0,518	6,331	0,214	3,406	0,290	4,430	0,295	4,273	0,325	4,649	0,335	2,776
b_childdensity	−0,052	−2,183	−0,039	−2,725	−0,038	−2,065	−0,043	−2,563	−0,045	−2,485	−0,063	−2,615
b_distwork	−0,142	−16,847	−0,076	−13,284	−0,099	−12,500	−0,097	−12,222	−0,095	−12,698	−0,075	−8,202
b_popyoung	0,018	10,319	0,014	10,701	0,015	9,768	0,017	11,288	0,018	11,452	0,014	7,533
b_rentratio	−1,227	−4,855	−1,133	−5,104	−1,003	−4,708	−0,960	−4,480	−1,051	−4,958	−0,796	−2,671
b_taxindex	−0,015	−4,087	−0,015	−5,487	−0,017	−5,153	−0,018	−5,519	−0,018	−5,806	−0,015	−3,711
b_timetoplatz	0,073	11,579	0,047	10,421	0,055	9,624	0,056	9,460	0,057	10,080	0,048	6,515

Figure 5.6: *Parameter estimation with the cluster sampling on CPCA (size=20)*

		Random			Cluster			Cluster Acpr		
	True values	Mean	Differences	Differences in absolute value	Mean	Differences	Differences in absolute value	Mean	Differences	Differences in absolute value
b_access	0,518	0,68	-0,162	0,162	0,292	-0,226	0,226	0,422	-0,096	0,096
b_childdensity	-0,052	-0,033	-0,019	0,019	-0,046	0,006	0,006	-0,043	0,009	0,009
b_distwork	-0,142	-0,075	-0,067	0,067	-0,089	0,053	0,053	-0,087	0,055	0,055
b_popyoung	0,018	0,014	0,004	0,004	0,016	-0,002	0,002	0,014	-0,004	0,004
b_rentratio	-1,227	-0,859	-0,368	0,368	-0,988	0,239	0,239	-0,930	0,297	0,297
b_taxindex	-0,015	-0,015	0	0	-0,016	-0,001	0,001	-0,012	0,003	0,003
b_timetoplatz	0,073	0,052	0,021	0,021	0,053	-0,020	0,020	0,047	-0,026	0,026
TOTAL				0,641			0,549			0,490

Figure 5.7: *Differences between mean of the parameters calculated on the reduced choice sets and the true values (size=20)*

	Random Sampling	Cluster Sampling (PCA)	Cluster Sampling (CPCA)
b_access	0,04500	0,05000	0,00227
b_childdensity	0,00000	0,00000	0,00011
b_distwork	0,00000	0,00000	0,00014
b_popyoung	0,00000	0,00000	0,00000
b_rentratio	0,01100	0,00500	0,01573
b_taxindex	0,00000	0,00000	0,00000
b_timetoplatz	0,00000	0,00000	0,00002
TOTAL	0,05600	0,05500	0,01827

Figure 5.8: *Variance of parameters across the 5 runs (size=20)*

A third table 5.9 shows instead the differences between the true values and all the estimated values. We don't insert here all the differences but we can see directly the sum of these differences and we can note how the CPCA cluster shows once again a lowest value.

	Random Sampling	Cluster Sampling (PCA)	Cluster Sampling (CPCA)
b_access	1,2311	1,0538	1,1287
b_childdensity	0,0957	0,0449	0,0550
b_distwork	0,3384	0,2793	0,2689
b_popyoung	0,0217	0,0172	0,0117
b_rentratio	1,8357	1,4844	1,1910
b_taxindex	0,0052	0,0169	0,0084
b_timetoplatz	0,1035	0,1299	0,1027
TOTAL	**3,6313**	**3,0264**	**2,7663**

Figure 5.9: *Total differences between true values and all the parameters computed on the reduced choice-sets (size=20)*

The computation of probability to be chosen for every alternative is another measure relative to the quality of the sampling. In this case we consider only one table in which we have the results for every sample and for all the sizes.

	Sample1	Sample2	Sample3	Sample4	Sample5	Sum	Var
Random	0,406609	0,503545	0,640277	0,677231	0,684295	2,911957	0,014961
Cluster PCA	0,375327	0,357611	0,366099	0,380373	0,366116	1,845526	0,000079
Cluster CPCA	0,314678	0,323959	0,32715	0,32373	0,329911	1,619427	0,000033

Figure 5.10: *Sum of the differences in absolute values between the probability for the alternatives, calculated with the true values and the values estimated on the reduced choice sets*

We see also in the figure 5.10 that the situation is better for the CPCA cluster technique both for the sum of the differences and for the variability. Last criteria to be evaluated is relative to the log-likelihood function. In this

case, we consider again what's happen for the size equal to 20. In the figure 5.11 we have simultaneously the bias, the variance and the total error.

	Bias	Variance	Total error
Random cluster	1420,256	32330,036	7802,513
PCA cluster	952,962	75730,023	4326,445
CPCA cluster	743,819	6948,974	3753,965

Figure 5.11: *Evaluation of ability to estimate overall Log-Likelihood function value (size=20)*

Here it's clear how the "CPCA cluster" give a big improvement to the estimation of the true log-likelihood function, and we can note how the variability across the 5 runs is very slow compared to the variability of the other two techniques. Also in this case the results are similar for the other sample sizes.
Sampling of alternatives can, as we said in past chapters, reduce computational time compared to using a full choice set, but the efficiency and the empirical accuracy of the estimated parameters is not always guaranteed. In this chapter we showed that the techniques we proposed ("CPCA cluster" and "PCA cluster" sampling) can improve the accuracy of sampling. Obviously, as with any numerical exercise, the usual cautions for generalizing the results apply also to this dissertation. There is certainly a need for more computational and empirical research on the topic of sampling of alternatives to draw more definitive conclusions. Anyway we think, looking at the results we obtained for the different sizes we chose, that, when the full choice set is to big to be used, the Cluster Sampling of Alternatives could be an useful technique to obtain good estimation of the parameters. In the classical random sampling, in fact, we dont know what kind of alternatives we select, so its possible that we could obtain all the alternatives with similar characteristics and so there could be some problems in the estimation. With the Cluster Sampling instead we obtain a choice set which reflects better the

full one.
As we said previously, the CPCA can be very useful also for the aggregation of the alternatives, and in next section we will show some results we obtained, also if the research about this topic must be study in more depth.

5.6 Aggregation of alternatives

The second kind analysis we show in this dissertation, once again based on CPCA with spatial constraint, is the Aggregation of the Alternatives. As we saw in the previous chapter, generally the aggregation of alternatives is done only according to geographical attributes. Sometimes instead the aggregation is conducted according to some particular variables observed on the elemental choices. In our proposal we join the two approaches. In this case the difference with the Cluster Sampling is that we don't stop the aggregation according to the classical criteria, but we decided, starting from 696 alternatives to obtain 35 aggregate alternatives. Our choice is justified considering the proportion for aggregation used in previous works (Hammadou et al. 2004) and, furthermore the number was adapted to the number of aggregate alternatives we could obtain carrying out the aggregation only according to geographic coordinates. This was necessary because in this way we could do a comparison between the two procedure utilized. In the application of these techniques an usual problem is the definition of the variables for the aggregate alternatives. In fact for some variables there are not problem mainly for spatial aggregation, because they are variables defined to a zonal level and so the values of elemental alternatives are the same also for aggregate alternatives. For other variables the solution we adopted was to consider, as done in other papers, the mean of the values observed on the elemental alternatives that constitute the aggregate one. Obviously this solution will cause some problems; another problem arised when we carried out the estimation. In fact we couldn't introduce in the utility function the

corrections we saw in the formula 3.27 relative to the heterogeneity of the elemental choices and to the number of disaggregate alternatives in the aggregate one. Anyway we introduce here the results we obtained applying the two techniques. In the tables 5.12 and 5.13 we can see the differences between the two procedures.

Name	Value	Std err	t-test		Robust Std err	Robust t-test	
b_access	0,6656	0,0937	7,1040		0,0960	6,9360	
b_childdensity	-0,0184	0,0307	-0,5998	*	0,0314	-0,5870	*
b_distwork	-0,1414	0,0071	-19,9727		0,0084	-16,9314	
b_popyoung	0,0011	0,0031	0,3563	*	0,0031	0,3647	*
b_rentratio	-0,4625	0,5426	-0,8524	*	0,3958	-1,1684	*
b_taxindex	-0,0075	0,0040	-1,8917	*	0,0035	-2,1456	
b_timetoplatz	0,0440	0,0066	6,6821		0,0074	5,9737	

Figure 5.12: *Parameter estimation after spatial aggregation*

Name	Value	Std err	t-test		Robust Std err	Robust t-test	
b_access	0,5219	0,1458	3,5802		0,1179	4,4263	
b_childdensity	0,0347	0,0656	0,5294	*	0,0537	0,6471	*
b_distwork	-0,0435	0,0174	-2,4962		0,0152	-2,8689	
b_popyoung	-0,0022	0,0055	-0,3916	*	0,0044	-0,4944	*
b_rentratio	-1,7690	0,2974	-5,9481		0,2435	-7,2660	
b_taxindex	-0,0496	0,0112	-4,4270		0,0084	-5,8819	
b_timetoplatz	0,0147	0,0124	1,1883	*	0,0093	1,5812	*

Figure 5.13: *Parameter estimation after "CPCA aggregation"*

In the two cases we see that there are three parameters not significant. The *childdensity* (measuring the density of children) and the *popyoung* (measuring the density of young people) have low values of the robust t-test in the

two cases, while the third parameter not significant is the *rentratio* (ratio between rent and income) for the spatial aggregation and the *timetoplatz* (time to go in the centre of the city) for the "CPCA aggregation". For other parameters the differences with the values calculated on the full choice set are minimal in both the cases. Anyway these results will be object of further studies to improve the quality of aggregation.

Appendix

S-PLUS CPCA code

```
   cpca <- function(mat1, mat2, mat3)
{
print("Data matrix")
print(mat1)
print("Spatial coordinates")
print(mat2)
print("Weights matrix")
print(mat3)
prossimityoperator
n <- nrow(mat1)
contiguitymatrix <- matrix(0, n, n)
for(i in 1:n) {
for(j in 1:n)
if(mat2[i] == mat2[j])
contiguitymatrix[i, j] = 1 }
diag(contiguitymatrix) <- 0
pmatrix <- contiguitymatrix %*% mat3
columnmarginal <- apply(pmatrix, 1, sum)
fmatrix <- diag(columnmarginal, nrow = nrow(mat1))
prossimityoperator <-fmatrix - pmatrix
#standardization
```

```
meanmat < − matrix(rep(apply(mat1, 2, mean), n), nrow = n, byrow = T)
variancemat < − matrix(rep(apply(mat1, 2, var), n), nrow = n, byrow =
T)/(n) * (n - 1)
standardizedmat < − (mat1 - meanmat)/(sqrt(n) * sqrt(variancemat))
#diagonalization
mat5 < − standardizedmat %*% solve(t(standardizedmat) %*% mat3 %*%
standardizedmat) %*% t(standardizedmat) %*% mat3 %*% prossimityoper-
ator %*% mat3 %*% standardizedmat %*% solve(t(standardizedmat) %*%
mat3 %*% standardizedmat) %*% t(standardizedmat)
components < −eigen(mat5)
eigenvalues < − as.numeric(components$values)
eigenvectors < − matrix(as.numeric(components$vectors), nrow = n, byrow
= F)
coordinates < − eigenvectors[, 1:56]
return(eigenvalues, coordinates)
}
```

S-PLUS code to build the reduced subset

```
subset_function(mat1, mat2, mat3, mat4, mat5, x, y, z)
{
n1 < − round(ncol(mat1)/696 * z)
n2 < − round(ncol(mat2)/696 * z)
n3 < − round(ncol(mat3)/696 * z)
n4 < −round(ncol(mat4)/696 * z)
n5 < − round(ncol(mat5)/696 * z)
id < − matrix(1:696, 696, 1)
choice < −matrix(1, 696, 1)
first < − matrix(0, 696, n1)
second < − matrix(0, 696, n2)
third < −matrix(0, 696, n3)
fourth < − matrix(0, 696, n4)
```

```
fifth < − matrix(0, 696, n5)
for(i in 1:696) {
first[i, ] < − sample(mat1[i, ], n1)
second[i, ] < − sample(mat2[i, ], n2)
third[i, ] < − sample(mat3[i, ], n3)
fourth[i, ] < − sample(mat4[i, ], n4)
fifth[i, ] < − sample(mat5[i, ], n5)
final < − cbind(first, second, third, fourth, fifth)
ordin < − final
for(i in 1:nrow(x))
for(j in 1:ncol(x)) {
if(x[i, j] == i)
ordin[i, 1] < − i
ordin[i, j] < − x[i, j]
}
for(i in 1:nrow(ordin))
for(j in 1:ncol(ordin)) {
if(ordin[i, j] == ordin[i, 1])
ordin[i, j] < − x[i, 1]
ordin[i, 1] < − ordin[i, 1]
}
nelm < − dim(y)[1] * dim(y)[2]
ngrp < − 11
nelmgrp < − 696
newmat1 < − ordin
elem < − nelmgrp
for(i in 2:ngrp) {
newmat1 < − cbind(newmat1, elem +ordin)
elem < − elem + nelmgrp
}
```

```
newmat2 < − t(as.matrix(newmat1[1, ]))
elemelem < − nelmgrp * ngrp
for(i in 2:dim(ordin)[1]) {
newmat2 < − rbind(newmat2, (elemelem + newmat1[i, ]))
elem < − elem + (nelmgrp * ngrp)
}
endmat < − matrix(c(t(y))[c(t(newmat2))], nrow = dim(y)[1], byrow = T)
finalsubset < − cbind(id,choice,endmat)
return(finalsubset)
}
```

Bibliography

Alonso, W. (1964), *Location and land use*, Harvard University Press.

Aluja, T. & Lebart, L. (1984), Local and partial principal component analysis and correspondence, *in* 'Compstat, Phisica Verlag', Vienna, pp. 113–118.

Anas, A. & Chu, C. (1984), 'Discrete choice models and the housing price and travel to work elasticities of location demand', *Journal of Urban Economics* **15**, 107–123.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers.

Autant-Bernard, C. (2005), Where do firms choose to locate their rd? a spatial conditional logit analysis on french data, Technical report, CREUSET University of St-Etienne.

Ben-Akiva, M. & Bierlaire, M. (1999), Discrete choice methods and their applications in short term travel decisions, *in* R. Hall, ed., 'The Handbook of transportation Science', Kluwer, Dordrecht, The Netherlands, pp. 5–33.

Ben-Akiva, M. & Bolduc, D. (1996), Multinomial probit with a logit kernel and a general parametric specification of the covariance structure, Massachusetts Institute of Technology.

Ben-Akiva, M., Bolduc, D. & Walker, J. (2001), Specification, identification and estimation of the logit kernel model, Technical report, Massachusetts Institute of Technology.

Ben-Akiva, M. & Bowman, J. (1998), 'Integration of an activity-based model system and a residential location model', *Urban Studies* **35**(7), 1131–1153.

Ben-Akiva, M., Fadden, D. M., Garling, T., Gopinath, D., Walker, J., Bolduc, D., Borsch-Supan, A., Delquié, P., Larichev, O., Morikawa, T., Polydoroulou, A. & Rao, V. (1999), 'Extended framework for modeling choice behavior', *Marketing letters* **10**(3), 187–203.

Ben-Akiva, M. & Lerman, S. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, Cambridge, Ma.

Ben-Akiva, M. & Morikawa, T. (1990), 'Estimation of travel deman models from multiple data sources', *Transportation and Traffic Theory* pp. 461–476.

Bentler, P. M. (1980), 'Multivariate analysis with latent variables', *Annual Review of Psychology* **31**, 419–456.

Bhat, C. (1998), 'Accomodating variations in responsiveness to level of service variables in travel mode choice models', *Transporttation Research A* **32**, 495–507.

Bhat, C. & Guo, J. (2004), 'Mixed spatially correlated logit model: Formulation and application to residential choice modelling', *Transportation Research B* **38**(2), 147–168.

Bierlaire, M. (2003), Biogeme: a free package for the estimation of discrete choice models, *in* '3rd Swiss transport Research Conference', Monte Verità, Ascona.

Bierlaire, M. (2005), 'An introduction to biogeme version 1.4', biogeme.epfl.ch.

Bierlaire, M. & Lucadamo, A. (2006), Sampling of alternatives using multidimensional analysis, *in* 'MTISD 06', Procida.

Burgle, M. (2006), Residential location choice model for the greater zurich area, *in* '6th Swiss Transport Research Conference', Monte Verità, Ascona.

Chattopadhyay, S. (2000), 'The effectiveness of mcfadden's nested logit model in valuing amenity improvement', *Regional Science and urban economics* **30**, 23–43.

Cornillon, P., Amenta, P. & Sabatier, R. (1999), Three-way data arrays with double neighbourhood relations as a tool to analyze a contiguity structure, *in* M. Vichi & O. Otiz, eds, 'Classificationa and data analysis. Theory and application', Springer Berlino, pp. 263–270.

Cornillon, P., Sabatier, R. & Chessel, D. (1993), Analyse d'un cube sosu double contrainte de voisinage., *in* 'The 49th session of International Statistical Insitute', Firenze, pp. 285–286.

D'Ambra, L. & Lauro, C. (1982), 'Analisi in componenti principali in rapporto a un sottospazio di riferimento', *Rivista di Statistica Applicata* **4**.

D'Ambra, L. & Lauro, N. (1992), 'Non symmetrical exploratory data analysis', *Statistica Applicata - Italian Journal of Statistics* **4**(4), 511–529.

D'Ambra, L. & Lucadamo, A. (2006), 'Constrained principal component analysis as an instrument to aggregate alternatives in logit models', working paper.

D'Ambra, L., Rodia, G. & Pagliara, F. (2005), Spatial dimension of choice models and zoning processes within a transportation system, *in* 'S.I.S. 2005 - Statistica a Ambiente', Messina.

Domencich, T. & McFadden, D. (1975), *Urban Travel Demand - A behavioral Analysis*, North Holland, Amsterdam.

Florax, R. & Rey, S. (1995), The impacts of misspecified spatial interaction in linear regression models: a meta analysis of simulation studies, *in* L. Anselin, R. J. M. Florax & S. Rey, eds, 'Advances in Spatial Econometrics: Methodology, Tools and Applications', Heidelberg: Springer.

Gabriel, S. A. & Rosenthal, S. S. (1989), 'Household location and race: estimation of a multinomial logit model', *The review of economic and statistics* **17**(2), 240–249.

Goetzke, F. (2003), Are travel demand forecasting models biased because of uncorrected spatial autocorrelation?, *in* 'North American Meeting of the Regional Science Association International'.

Griffith, D. A. (1996), Some guidelines for specifying the geographic weights matrix contained in spatial statistical models, *in* 'Practical Handbook of Spatial Statistics', S. L: Arlinghaus, Boca Raton: CRC.

Guo, J. Y. (2004), Addressing Spatial Complexities in Residential Location Choice Models, PhD thesis, University of Texas, Austin.

Hammadou, H., Thomas, I., Tindemans, H., Witlox, F., Hofstraeten, D. V. & Verhetsel, A. (2004), How to incorporate the spatial dimension in destination choice models?, *in* 'Convergence et disparités régionales au sein de l'espace européen. Les politiques régionales à l'épreuve des faits', Bruxelles.

Hausman, J. & McFadden, D. (1984), 'Specification tests for the multinomial logit model', *Econometrica* **46**, 403–426.

Jolliffe, I. (2002), *Principal Component Analysis*, Springer.

Kanaroglou, P. S. & Ferguson, M. R. (1998), 'The aggregated spatial choice model vs multinomial logit: an empirical comparison using migration data', *The Canadian Geographer* **42**(3), 218–231.

Lancaster, K. (1966), 'A new approach to consumer theory', *Journal of Political Economy* **74**, 132–157.

Lebart, L. (1969), Analyse statistique de la contiguité, Technical Report 28, Publcation de l'Institut de Statistiques de l'Université de Paris. 81-112.

Lebart, L., Morineau, A. & Warwick, K. (1984), *Multivariate descriptive statistical analysis*, J. Wiley, New York.

LeFoll, Y. (1982), 'Ponderation des distances en analyse factorielle', *Statistiques et Analyse des Données* **7**, 13–31.

Lerman, S. R. (1983), Random utility models of spatial choice, *in* 'Optimization and discrete choice in urban system', B. G: Hutchinson and P. Nijkamp and M. Batty.

Levine, J. (1998), 'Rethinking accessibility and jobs-housing balance', *Journal of American Planning Association* **64**(2), 133–149.

Luce, R. D. & Tukey, J. W. (1964), 'Simultaneous conjoint measurement: a new type of fundemental measurement', *Journal of Matematical Psychology* **1**, 1–27.

Manski, C. (1973), The Analysis of Qualitative Choice, PhD thesis, Department of Economics, MIT, Cambridge, Mass.

Marschak, J. (1960), Binary choice constraints on random utility indications, *in* K. Arrow, ed., 'Stanford Symposium on Mathematical Methods in the Social Sciences', Stanford University Press, Stanford, pp. 312–329.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka, ed., 'Frontiers in Econometrics', New York: Academic Press, pp. 105–142.

McFadden, D. (1978), Modelling the choice of residential location, *in* A. Karlqvist, L. Lindqvust, F. Snickars & J. Weibull, eds, 'Spatial Interaction Theory and Planning Models', North-Holland, Amsterdam.

McFadden, D. (1984), Econometric analysis of qualitative response models, *in* Z. Friliches & M. Intriligator, eds, 'Handbook of Econometrics II', Elsevier Science Publishers.

McFadden, D. & Train, K. (2000), 'Mixed mnl models for discrete response', *Journal of Applied Econometrics* **15**(5), 447–470.

McFadden, D., Tye, W. & Train, K. (1977), 'An application of diagnostic tests for the irrelevant alternatives property of the multinomial logit model', *Transportation Research Record* **637**, 39–46.

Meot, A., Chessel, D. & Sabatier, R. (1993), Opérateurs de voisinage et analyse des données spatio-temporelles, *in* Asselain, ed., 'Biometrie et Environnement', Masson, Paris.

Miyamoto, K., Vichiensan, V., Shimomura, N. & Paez, A. (2004), 'Discrete choice model with structuralized spatial effects for location analysis', *Transportation Research Record, Travel Demand and Land Use* **1898**, 183–190.

Mohammadian, A. & Kanaroglou, P. (2003), Applications of spatial multinomial logit model to transportation planning, *in* 'Conference Paper Moving through nest: The physical and social dimensions of travel', Lucerne.

Mom, A. (1998), 'Eigenstructure of distance matrices with an equal distance subset', *Linear Algebra and its Applications* **280**, 245–251.

Monestiez, P. (1978), Méthodes de classification automatique sous contraintes spatiales, *in* 'Biométrie et Ecologie', .M. Legay and R. Tomassone, pp. 367–379.

Nelson, G., Pinto, A. D., Harris, V. & Stone, S. (2004), 'Land use and road improvements: A spatial perspective', *International Regional Science Review* **27**, 297–325.

Nerella, S. & Bhat, C. (2004), 'A numerical analysis of the effect of sampling of alternatives in discrete choice models', TRB 2004: For presentation and publication.

Parsons, G. & Needelman, S. (1992), 'Site aggregation in a random utility model of recreation', *Land Economics* **68**(4), 418–433.

Pellegrini, P. & Fotheringham, A. (2002), 'Modeling spatial choice: a review and synthesis in a migration context', *Progress in Human Geography* **26**(4), 487–510.

Quigley, J. (1976), 'Housing demand in the short run: An analysis of polytomous choice', *Explorations in Economic Research* **3**(1), 76–102.

Revelt, D. & Train, K. (1998), 'Mixed logit with repeated choices: Households' choice of appliance efficiency level', *Review of Economics and Statistics* **80**(4), 647–657.

Small, K. & Hsiao, C. (1982), Multinomial logit specification tests., Technical report, Depatment of Economics Princeton University, Princeton, N. J.

Thioulouse, J., Chessel, D. & Champely, S. (1995), 'Multivariate analysis of spatial patterns: a unified approach to local and global structures', *Environmental and Ecological Statistics* **2**, 1–18.

Thurstone, L. (1927), 'A law of comparative judgement', *Psychological Review* **34**, 273–286.

Tobler, W. (1970), 'A computer modeling simulating urban growth in the detroit region', *Economic Geography* **46**(2), 234–240.

Tobler, W. (1991), Frame independent spatial analysis, *in* M. Goodchild & S. Gopal, eds, 'Accuracy of Spatial Database', Taylor and Francis, New York, pp. 115–122.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press.

Varian, H. R. (1992), *Microeconomic Analysis, Third edition*, Norton Company, New York.

Wartenberg, D. (1985), 'Multivariate spatial correlation: a method for explanatory geographical analysis', *Geographical Analysis* **17**, 263–283.

Wong, D. (1996), Aggregation effects on geo-referenced data, *in* 'Practical Handbook of Spatial Statistics', S. Arlinghaus CRC press, Boca Raton Florida, chapter 5.