Università degli Studi di Genova
DISEFIN – Series of Economic Working Papers
16126 Genova – via vivaldi 5 – Fax +39 010 209 5223

A Multivariate Analysis Of The Space Syntax Output For The Definition Of Strata In Street Security Surveys

Enrico di Bella, Luca Persico, Matteo Corsi

wp n. 5
September 2011

"DISEFIN Working Papers on line"
series of economic working papers
published online by

Research Doctorate in
Public Economics
Ph.D School in New technologies and Social Sciences
University of Genoa

Founder:
Amedeo Fossati

Editor-in-Chief:
Marcello Montefiori

Web site:
http://www.disefin.unige.it/

A Multivariate Analysis Of The Space Syntax Output For The Definition Of Strata In Street Security Surveys

Enrico di Bella
University of Genoa, Italy

Luca Persico
University of Genoa, Italy

Matteo Corsi
Ass. Kallipolis Trieste, Italy

Abstract

Although the connection between crime and urban layout is generally evident, surveys inquiring that relationship are often facing two different problems: areas with high criminality are often inhabited by partially elusive populations (being stowaways) and the urban structure (e.g. length and width of streets) quickly changes even after a few corners. In this work a combination of two techniques well known in their specific field is proposed to define a simple two-stages sampling design. Space Syntax is a set of measurements which are done on the topographic maps of a town with the division of all the roads into segments, called axes. Using multivariate techniques, these axes can be classified on the basis of a homogeneity criterion obtaining the strata for a two-stages sampling design.

Keywords: *Factor Analysis, Geodetic networks, Street security surveys, Space Syntax, Urban axes*

Jel Classification: *C83, O18*

1. Introduction

In the last 20 years, most European cities have undergone vigorous renovation programmes aimed substantially (when not exclusively) at increasing safety and bolstering crime prevention. Official acts of the EU have been reflecting this tendency at least since 1994, with the inauguration of the "URBAN I" Programme. A debate on the role of urban planning on crime and crime prevention, however, dates back at least to the 1950s and 1960s when large urban renewal projects were put in place in the United States and their consequences spurred the well know and very influential argument between New York City planner Robert Moses and American/Canadian urbanist Jane Jacobs (Jacobs, 1961).

Such an effort reflects the fact that, despite the vast number of conflicting theories on the causes of crime, there is widespread consensus on the idea that crime is a mostly urban phenomenon (Durkheim, 1897 and Weber, 1958) and that its roots have to be looked for somewhere in between the social structure of urban communities and the structure of the built environment they inhabit.

Each of the main international organizations that in some form deal with the subject of crime (from the World Bank to different branches of the UN galaxy) have specific workgroups, programmes, projects and offices dedicated to urban crime, creating a framework that is intended to give an impulse to urban crime-prevention research and policies. Most notably, UNODC (United Nations Office on Drugs and Crime), whose "Commission on Crime Prevention and Criminal Justice" has urban crime among its mandated priority areas, UN-Habitat (United Nations Human Settlements Program) throughout its "Safer Cities Programme" and several country projects of the World Bank (in particular in Latin America).

Accordingly, "Aménagement urbain et sureté: les issues de deux cas italiens et l'étude d'un nouveau modèle d'analyse orienté à

l'action" is a project funded by Plan Urbanisme Construction Architecture, an interdepartmental body of the French Republic, via an international research programme on urban safety. The project is designed to investigate, in five circumscribed areas of two major Italian cities, the existing correlations between urban form, pedestrian movement and safety, in an attempt to estimate the impact of urban spatial configuration on natural surveillance (Jacobs, 1961) and, in turn, that of natural surveillance on actual crime figures.

This goal will be pursued by comparing and complementing the mentioned impacts with more traditional measures of urban liveability and safety.

The expected outcome of the research is a non-contextual model of urban analysis that allows researchers and governmental authorities to plan renovation programmes consisting of a rational and motivated combination of welfare/community development, crime prevention through environmental design and spatial and configuration/natural surveillance management as a bridge between them.

Three are the main elements that the project assumes as measurable: the victimization rate at street level through victimization surveys, the pedestrian flow at street level through on-site pedestrian count, and the configurational properties of the urban grid through Space Syntax Analysis. In fact, each element is measured at more than a street level, as the actual unit of measurement is the "axial line", a space syntax analysis concept that is introduced in the next paragraph and that frequently covers a scale smaller than the whole street.

Because of the variety and wideness of the various topics involved in this project, in this paper the authors will focus exclusively on the usage of the Space Syntax Analysis output to define a sampling design that will be used to set up crime victimization surveys.

2. The space syntax analysis

Space Syntax Analysis is a technique that measures the configurational properties of an urban grid through the methodological features of graph theory (Hillier & Hanson, 1984). In graph theory, a graph $\mathcal{G}$ is a network consisting of two sets of information: a set of nodes, $\mathcal{N} = \{n_1, n_2, \ldots n_N\}$ and a set of lines $\mathcal{L} = \{l_1, l_2, \ldots l_L\}$, each connecting two nodes. As a consequence, in an undirected graph, each line can be unambiguously identified by the unordered pair of distinct nodes that it connects, therefore:

$$l_k = (n_i, n_j) = (n_j, n_i) \tag{1}$$

The urban layout of a city and, specifically, its system of roads and public spaces intertwining the buildings, can be thought as a non directional network.

That network can be represented as a graph and then analyzed either adopting the standard operations of graph theory or formulating new ones, specifically designed to analyze space and human behaviour in space.

The formal representation of an urban layout through a graph is done by re-drawing a bi-dimensional urban grid as an axial map, a network consisting of the fewest, longest straight walkable lines (axial lines) that cover the whole grid, and their intersections (Figure 1). The axial map is then analyzed as a graph where, counter intuitively, each axial line is assumed as a node or vertex of the graph and each intersection between two axial lines as an edge of the graph.

In Graph Theory, nodes represent actors and lines represent ties between the actors (Wasserman an Faust, 1994). In Space Syntax, spaces are the actors and intersections are their ties.

Since a graph is, by default, the way to represent mathematically an axial map, it goes without saying that, from this point on,

"node" and "axial line" are going to be used somewhat interchangeably. Figure 2 shows the relationship between an axial map and its corresponding graph.

With this procedure, Space Syntax Analysis describes in measurable terms the empty spaces of a map and their configuration. However, a topological description of the empty spaces of a map is, by definition, one that ignores some of their measurable features, like their linear length and metric properties, in favour of providing a syntax for expressing their position in a way that is not sensitive to continuous deformations.

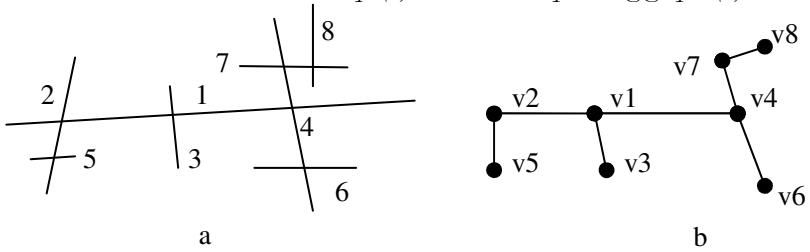
The reason for expressing positions and describing space using this approach was the one of exposing the limited number of actual models of space configurations that can be found in human settlements when continuous deformations are not considered, explicating the social and cultural implications of those models.

In that form, Space Syntax was initially conceived by Prof. Bill Hillier and Julienne Hanson at The Bartlett, University College of London, in

FIGURE 1. – *Axial map overlapping the cartographic map.*

the 1970s and in the 1980s, as a theory to analyze small environments and their configurational features. The theory received substantial impulse by the increase in the computational capacity of computers that allowed its use on larger graphs and made it possible to experimentally test it on reasonably vast built environments (Cities as vast as London and Atlanta have been represented through axial maps and can be analyzed with the appropriate software and thanks to up-to-date hardware in a few seconds of computation time). The implications of Space Syntax Analysis, however, became even more impressive when a substantial relation was found between individual behaviour in space, cognition of space and movement through it, and the configuration of space expressed in topological terms. Since then, space syntax has been studied as an increasingly convincing predictor of pedestrian movement.

FIGURE 2. – *Axial map (a) and its corresponding graph (b)*



3.Measures in space syntax analysis

Most of the measures provided by Space Syntax Analysis represent either a way to express the degree to which a node is connected to its immediate surroundings or its centrality with respect of the whole graph or part of it. Table 1 summarizes the Space Syntax measures analyzed and used in this work.

Connectedness is usually referred to a node and the nodes it

is directly linked to through a vertex. This makes it (and each of the different ways to measure it) a local information, i.e. an information regarding specific parts of the graph.

Centrality extends its meaning beyond direct connections and largely depends on a definition of distance that is peculiar to Graph Theory. Length is not intended as linear length, but as a count of the intermediate nodes that stand on the shortest path between *i* and *j*. Consequently, different measures of centrality can express local or global information.

A measure of centrality can easily be affected by the global size of the system and, therefore, two measures of centrality regarding two different graphs may not be comparable if the two graphs are different in size. Because of that, various measures of centrality for a node are normalized in some way and the procedure that will be followed throughout this paper is that of the *D*-Value (Kruger, 1989). Normalization through the *D*-Value is obtained comparing a centrality measure of the *i*-th node of a graph with *n* nodes with the centrality measure we would get if the node was at the root of a different graph, again consisting of *n* nodes but standardised in a diamond shape. According to (Kruger, 1989; Teklenburg et al. 1993 and Hillier and Hanson, 1983), a diamond-shaped graph is a particular form of justified graph in which a node (the root) is put at the base and then all nodes at depth 1 are aligned horizontally above it, all nodes at depth 2 from the root are aligned above those at depth 1 and so on until all levels of depth are accounted for. In a diamond shape there are k nodes at mean depth level, k/2 nodes at one level above and below and k/4 at two levels above and below up to the point where we have a single node at the deepest level and one at the shallowest (the root itself). In such a graph, the depths from a root are, approximately, normally distributed (Kruger, 1989; Hillier and Hanson, 1983); therefore, comparing the RA value of its root to that of a node in a graph with the same number of nodes is a way to compare a normal distribution with the actual distribution. In

such a graph it is possible to demonstrate (Hillier and Hanson, 1983) that the D-Value is:

$$D_n = \frac{2\left\{n\left[\log_2\left(\frac{n+2}{3}\right)-1\right]+1\right\}}{(n-1)(n-2)} \qquad (2)$$

being n the number of nodes in the axial map.

TABLE 1A. – Formulas, meanings and descriptions of the space syntax output variables used in the subsequent analyses (continues).

Variable	Formula, meaning and description of the variable	
Connectivity	$C_i = k_i$	Number of axes connected to the i-th axis.
	Measures how much an axis is directly connected to the others. Axes with high C_i make it easier to pass through the areas.	
Control	$CTRL_i = \sum_{j=1}^{k_i} \frac{1}{C_j}$	Sum of the inverse values of C_i for the k_i axes connected to the i-th axis.
	Strategic relevance of an axis as the main or only connection to the whole system for the constellation of axes directly linked to it. *CTRLi* is maximized if the axis is connected only with dead ends.	
Controllability	$AB_i = \frac{k_i}{w+1}$	Ratio between connectivity and the $w + 1$ axes at a distance $d_{ij} \leq 2$.
	High values indicate axes that are easy to dominate (visually, for example) from a nearby vantage point.	
Mean Depth	$MD_i = \frac{1}{n-1}\sum_{j=1}^{n-1} d_{ij},\ i \neq j$	M distance of the i-th axis from all the other $n - 1$.
	Basic centrality measure: it accounts for the distance between each axis and all the others, with the shallowest axis being the closest to all the others and the deepest being the farthest one. In a city with a perfectly circular axial map, the shallowest axis would be close to the centre and the deepest would be on the perimeter.	

TABLE 1B. – Formulas, meanings and descriptions of the space syntax output variables used in the subsequent analyses (continued).

Variable	Formula, meaning and description of the variable	
Mean Depth R2	$_{R2}MD_i = \frac{1}{w-1}\sum_{j=1}^{w-1} d_{ij}, \quad i \neq j$	Average distance of the i-th axis from the other $w-1$ axes at a distance $d_{ij} \leq 2$.
	The meaning is the same as above, but those properties are referred to just a part of the axial map, the one standing at a topological distance of 2 o less from the examined axis.	
Relative Asymmetry	$RA_i = \frac{2(MD_i - 1)}{n-2}$	Normalized value of MD_i being $\min(MD_i) = 1$ and $\text{Max}(MD_i) = n/2$.
	RA expresses the centrality of an axis comparing its actual Mean Depth with the theoretical highest and lowest values that Mean Depth could have in the given graph. Compared to Mean Depth alone, Relativized Asymmetry is a normalization between 0 and 1.	
Relative Asymmetry R2	$_{R2}RA_i = \frac{2(_{R2}MD_i - 1)}{n-2}$	Normalized value of $_{R2}MD_i$ being $\min(_{R2}MD_i) = 1$ and $\text{Max}(_{R2}MD_i) = n/2$.
	Same as RA_i, but focused on the local structure of the axial map since Mean Depth is replaced with Mean Depth R2.	
Real Relativized Asymmetry	$RRA_i = \frac{RA_i}{D_i}$	Measurement of RA_i relatively to D_i. If $RRA_i = 1$the graph is "diamond".
	RRA is a normalized measure since it is calculated as RA normalized through the D-Value. This makes RRA a centrality measure that is independent from the size of the graph, and RRA values from different graphs are hence comparable.	
Real Relativized Asymmetry R2	$_{R2}RRA_i = \frac{_{R2}RA_i}{D_i}$	Measurement of RA_i relatively to D_i. If $RRA_i = 1$ the graph is a "diamond".
	Again, same as above, but within the perimeter of the axes at a topological distance of 2 or less from the axis we are analyzing	

TABLE 1C. – Formulas, meanings and descriptions of the space syntax output variables used in the subsequent analyses (continued).

Variable	Formula, meaning and description of the variable	
Integration	$INT_i = \frac{1}{RRA_i} = \frac{D_i}{RA_i}$	Inverse of RRA_i.
	Standard global measure of centrality to be used with RRA. High levels of Integration define the small areas, close to the geometric centre of the axial map, made of both long and short axes.	
Integration R2	${}_{R2}INT_i = \frac{1}{{}_{R2}RRA_i} = \frac{D_i}{{}_{R2}RA_i}$	Inverse of ${}_{R2}RRA_i$.
	Centrality measure for axes at a distance of 2 or less from the i-th.	
Choice	$C_i = \frac{\sigma_{s,t}(i)}{\sigma_{s,t}}$	Geodetic paths between each couple of nodes going through (i) over total paths.
	Global measure of centrality (see INT_i) done on highly hierarchical measurements, with most axes having very low values and few of them, long axes that constitute the backbone of the urban fabric, with values much higher than the average.	
Choice R2	${}_{R2}C_i = \frac{\sigma_{s,t}(i)}{\sigma_{s,t}} \forall s,t \;\; \exists\, d_{si} \le 2, d_{ti} \le 2$	Like Choice, but with each node of the (s,t) couples at distance 2 or less from (i).
	As usual with R2 values, Choice R2 has the same meaning as Choice, but it only refers to a part of the city that stands within 2 steps of topological distance from the axis we are analyzing.	
Node Count R2	n_i	Number of nodes at $d_{ij} \le 2$ connected to the i-th axis.
	Local measure of connection for an axis.	
Length	l_i	Metric length of the i-th axis.
	The length of an axis is a measure of its role inside the map.	

4. The sampling design and axes grouping

Sampling techniques are very numerous and differ from one another according to the applicatory contexts and the information available before the survey. Sampling requires a not necessarily nominative list of all the units in the population. In the present survey a partially difficult-to-sample population has to be faced as an exhaustive and accurate list of the units to be sampled does not exists. Although it is not possible to suppose a relevant number of homeless, in the areas analyzed (Porta Palazzo and Lingotto in Turin) the electoral list of voters is not a complete list of the residents as there is a relevant percentage of people living in the area but not therein officially resident or being illegal immigrant. For this reason the "classical" stratified sampling methodology based on gender and age classes can't be easily applied and the use of a multiphase sampling is suggested.

Moreover, the topic of the study is the definition of the relationship between the perception of safety that resident have and the urban characteristics or the axis in which they are living and, therefore, the focus should be on the structure of axes more than the composition of residents in the axes. The procedure applied, inspired by, but not exactly, an areal sampling, is defined through a few steps:

1. generation of the axes decomposition of the urban layout according to the space syntax approach and measurement of the space syntax variables specified in paragraph 3;
2. factorial analysis to resume the space syntax output variable into a few factors with an urban meaning;
3. clustering the axes using the two factors found in step 2 using a *k*-means approach obtaining the fist-stage units for the sampling;

4. definition of the number of axes to be sampled in the clusters identified proportionally to the population living in those axes over the total;
5. random selection of one building per sampled axis;
6. preliminary analysis and eventual integration of the first sample according to strata variances estimates.

The definition of cohesive subgroups (subsets of actors among whom there are relatively strong, direct, intense, frequent or positive ties) is quite common in the Social Networks Analysis (Wasserman an Faust, 1994). Many authors have discussed the role of social cohesion in social explanations and theories (Burt, 1984; Collins 1988; Erickson 1988; Friedkin 1984) but the literature about cohesive groups in geodetic networks is quite poor. Obviously the measurements used by the various authors to define cohesive subgroups are various but the methodologies used to create the clusters of actors in the network represent proximities among them are the classical statistical multivariate techniques: Principal Components and Factor Analysis (e.g. Bock and Husain, 1952; MacRae, 1960; Wright and Evitts, 1961), Multidimensional Scaling (e.g. Laumann and Pappi, 1973; Freeman, Romney and Freeman, 1987; Arabie 1977; Caldeira 1988), Clustering Techniques (Wille, 1984). In this work the Multidimensional Approach is not taken into account as Factor Analysis gave good results with a clear interpretation of the factors. Future works could analyze the effectiveness of the two different approaches.

4.1 Principal Components Analysis

Given a set of data, principal components analysis looks for a few linear combinations that can be used to summarize the data, losing in the process as little information as possible. As a first objective, principal component analysis seeks the Standardized Linear Combination of the original variables that has maximal variance.

If $\mathbf{x}$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then the principal component transformation is given by:

$$\mathbf{x} \rightarrow \mathbf{y} = \boldsymbol{\Gamma}'(\mathbf{x} - \boldsymbol{\mu}) \tag{3}$$

where $\boldsymbol{\Gamma}$ is orthogonal, $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{\Lambda}$ is diagonal and the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. The strict positivity of the eigenvalues λ_i is guaranteed if $\boldsymbol{\Sigma}$ is positive definite. The representation of $\boldsymbol{\Sigma}$ follows the Jordan Decomposition Theorem (Mardia *et al.*, 1979, p. 469). The *i*-th principal component of $\mathbf{x}$ is the *i*-th element of the vector $\mathbf{y}$, namely as

$$y_i = \gamma'_{(i)}(\mathbf{x} - \boldsymbol{\mu}) \tag{4}$$

where $\gamma'_{(i)}$ is the *i*-th column of $\boldsymbol{\Gamma}$ and may be called the *i*-th vector of principal component loadings. The function y_i is, then, the *i*-th principal component of $\mathbf{x}$.

It is easy to prove (Mardia et al, 1979, p. 215) the following theorem:

THEOREM 1 (*Properties of principal components*).
If $E(\mathbf{x}) = \boldsymbol{\mu}$, $Var(\mathbf{x}) = \boldsymbol{\Sigma}$ *and* $\mathbf{y}$ *is as defined in (3), then:*

1. $E(y_i) = 0$;
2. $Var(y_i) = \lambda_i$;
3. $Cov(y_i, y_j) = 0$, $i \neq j$
4. $Var(y_1) \geq Var(y_2) \geq \ldots \geq Var(y_p) \geq 0$
5. $\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} \lambda_i = tr\boldsymbol{\Sigma}$

4.2 Models for Factor Analysis

The basic idea underlying factor analysis (e.g., Press, 1972) is that p observed random variables X, can be expressed, except for n error term, as linear functions of m ($< p$) hypothetical variables or common factors, i.e. if $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ are the variables and $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m)$ the factors then:

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e} \tag{4}$$

where $\mathbf{\Lambda}$ is the matrix of the so called factor loadings λ_{ij} $i = 1, \ldots, p, j = 1, \ldots, m$.

As stated by Joliffe (1986), both Principal Component Analysis and Factor Analysis aim to reduce the dimensionality of a set of data, but the approaches used to do so are different for the two techniques.

Principal component analysis has been extensively used as part of Factor Analysis but there is not any explicit model for the former. The Factor Analysis requires six basic assumptions:

1. $E[\mathbf{e}] = \mathbf{0}$
2. $E[\mathbf{f}] = \mathbf{0}$
3. $E[\mathbf{x}] = \mathbf{0}$
4. $E[\mathbf{ee}'] = \mathbf{\Psi}$ (diagonal)
5. $E[\mathbf{fe}']=\mathbf{0}$ (a matrix of zeros)
6. $E[\mathbf{ff}]=\mathbf{I}_m$ (an identity matrix)

Whereas assumptions 1-3 are standard and convenient assumptions made in most statistical models without loss of generality, assumptions 4 and 5 are fundamental. The last assumption can be relaxed so that the common factors may be correlated (oblique) rather than uncorrelated (orthogonal). Many techniques in factor analysis have been developed for finding orthogonal factors, but some authors (e.g., Cattell, 1978) argue that oblique factors are almost always necessary to get a correct factor structure. Usually the assumption of multivariate normality is made but, like in PCA, many of the results do not depend on specific distributional as-

sumptions. Some restrictions are generally necessary on $\mathbf{\Lambda}$, because without any restrictions there will be a multiplicity of possible $\mathbf{\Lambda}$ which give equally "good" solutions. Because of this indeterminacy, estimation of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ precedes in two stages:

1. some restrictions are placed on $\mathbf{\Lambda}$ in order to find an initial solution (for example through PCA);
2. other solutions are found by rotation of $\mathbf{\Lambda}$ multiplying by an orthogonal matrix T and the "best" one is chosen according to some particular criterion.

Usually, it is not possible to define the best rotation criterion and many of them should be explored but all of them aim to obtain a rotated structure of Λ with most elements "close to zero" or "far from zero" with as few as possible elements taking intermediate values. The great advantage of rotation is that it simplifies the factor loadings or rotated PC coefficients, which can help in interpreting the factors or rotated PCs.

As mentioned earlier a major distinction between factor analysis and PCA is that there is the model (3) underlying factor analysis but no such model in PCA. Both techniques can be thought of as trying to represent some aspects of the covariance matrix Σ as well as possible, but PCA concentrates on the diagonal elements, whereas in the factor analysis the interest is the off-diagonal elements. Therefore, if a PCA is performed on n independent variables, there will be a PC corresponding to each such variable whereas a common factor in factor analysis must contribute to at least two of the variables, so it is not possible to have a "single variable" common factor. Thus, for a given dataset, the number of factors required for an adequate factor model will be no larger, and may be strictly smaller, than the number of PCs required to account for most variation in the data.

Despite of these differences, PCA and factor analysis both have the aim of reducing the dimensionality of a vector of random variables and the use of PCs to find initial factor loadings, though having no firm justification in the theory, will often not be misleading in practice.

For all the aforesaid reasons the use of principal components analysis and factors rotation in this context is justified and the results are reported in the subsequent paragraphs.

4.3 Cluster Analysis

The term "clustering" includes a collection of techniques that are used to group multidimensional entities according to various criteria of their degree of homogeneity and heterogeneity. The problem of clustering N data vectors in a p-dimensional space into k clusters can be handled in many ways (Everitt, 1974) and the most appropriate technique to use depends upon the problem. In this work we use a Non Hierarchical Clustering (k-means clustering) which is based upon the criterion of minimization of the variance within the clusters and which can be thought as an "ANOVA in reverse" as the methodology moves objects (e.g., cases) in and out of groups (clusters) to get the most significant ANOVA results. Usually, as the result of a k-means clustering analysis, we would examine the means for each cluster on each dimension to assess how distinct our k clusters are. Ideally, we would obtain very different means for most, if not all, dimensions used in the analysis.

4.4 Sampling Design

In most sampling problems the population can be regarded as being composed of a set of groups and elements (e.g. Kalton, 1989, Fabbris, 1989, Diana and Salvan, 1987, Cicchitelli, 1992, Gambini, 2009). One sampling use for such groups is to treat them as strata so that separate samples are selected from each group. If all the elements in the second-stage level are sampled, the method

is known as cluster sampling whereas if only a sample of elements is taken from each selected cluster, the method is known as two-stage sampling. Often a hierarchy of clusters is used: first some large clusters are selected, next some smaller clusters within the selected large clusters and so on until finally elements are selected within the final-stage cluster. The sampling procedure above specified can be expressed as a multi-stage sampling (or a cluster sampling if all the units in a building are sampled) in which the units sampled in the last stage are all the units of the last stage itself: the universe is divided into a number of first-stage (or primary sampling) units (axes), which are sampled; then the selected first-stage units are sub-divided into a number of smaller second-stage (or secondary sampling) units (buildings). The N units of the population are divided into H subpopulations (axes groups) and h axes are sampled from H, n_j ($j = 1, ..., H$) buildings are sampled from each of the h stages sampled (generally $n_i \neq n_j \quad i \neq j$), for a total of n units. In particular it is assumed to sample one building per axis.

Although strata and clusters are both groupings of elements, they serve entirely different sampling purposes. Since strata are represented in the sample, it is advantageous if they are internally homogeneous in the survey variables. On the other hand, with only a sample of clusters being sampled, the ones selected need to represents the ones unselected; this is best done when the clusters are as internally heterogeneous in the survey variables as possible. Proportionate stratification is used to achieve gains in precision but, generally, cluster sampling leads to a loss in precision compared with a simple random sample of the same size. The reasons why in this work this procedure is suggested are the following:

1. clustering is generated using objective measurements (although, some inconsistencies of Space Syntax have been reported by Ratti, 2004);

2. the cluster sampling procedure reduces the cost of the survey in comparison to the same SRS of the same size;
3. the list of the units composing the population is not available.

An aspect relevant for the analysis of the sample design is whether the clusters are equal in size or not. Generally the natural groupings that the sampler takes advantage of to serve as clusters almost always vary in size, often in a major way. In the following the difficulties this variation in size creates will be explained and methods by which they can be overcome will be described.

In general, with a two-stage design, the probability of element B in cluster A ($A \cap B$) appearing in the sample is:

$$P(A \cap B) = P(A)P(B|A) \tag{5}$$

where $P(A)$ is the probability of cluster A to be selected and $P(B|A)$ the probability of cluster B to be chosen at the second stage, given that cluster A was selected at the first stage. This equation, which can be extended to cover more sampling stages when necessary, is sometimes known as the selection equation. Consequently, if the selection of first-stage and second-stage clusters are purely random, probability (4) is strictly connected to the number of units in the clusters. Moreover, if all the units are equally relevant for the survey, the probability of selection of one cluster is supposed to be proportional to cluster size.

In practice, this "Probability proportional to size" sampling as described is seldom possible, because the true sizes of the sampling units are usually unknown. Often, however, good estimates are available from a recent census or some other source. For example, it is possible, from Local Administration data, to have the list of the regular residents in the areas of interest. Illegal immigrants and, more generally, elusive units can be assumed to be equally

distributed in the areas analyzed as they are very limited in surface and number of axes.

5. The Turin case

Turin (Torino) with almost one million inhabitants is the forth Italian city after Rome, Milan and Naples. Founded as a Roman military camp (*Castra Taurinorum*) in the first century BC, the city developed respecting the typical Roman street grid core as it can still be seen in the modern city.

The two areas of interest Porta Palazzo (PP) and Lingotto (LI) are topologically extremely different (Figure 3). The "Porta Palazzo" quarter takes its name from the homonymous city gate which had been built in 1701 in the northern part of Turin. This area, initially a suburban area outside the city walls, became an integrant part of Turin in 1800 as a decision of Napoleon. In the subsequent years it became a commercially relevant quarter with an important market (today in this quarter the biggest not covered European market takes place) and a generalized overpopulation (actually approx. fifteen thousand residents) due to massive immigration from southern Italy and non-EU countries with the successive urban decay. In recent years, and in particular after 1996, a strong requalification process is taking place in the area in the framework of the "The Gate-living not leaving" project financed by the European Union and various national institutions. The "Lingotto" quarter takes its name from the biggest farm active in this area when it was a rural burg. When, in 1915, Fiat decided to establish a factory in this area the urban growth of this area dramatically boosted making of Lingotto a big and highly populated working-class quarter (approx. fifty thousand residents). Nowadays, the Porta Palazzo and Lingotto areas count respectively 17.715 and 20.359 inhabitants (resident people). The demographic composition of the two populations, as we can see in Table 2 and Table 3, presents some specific features. Porta Palazzo area is characterised by an higher

percentage of not Italian residents corresponding to about 33% of the total population (Lingotto presents only 10% of not Italian residents). The age distribution is instead very similar in the two areas, about the 70% of the total residents are sixty or less. As it can be seen in Figure 3, the different percentage of people coming from non-UE countries between the two areas, determines a highly different population structure. In particular, the different percentage distribution between the two area of interest is due to the presence, in the Porta Palazzo area, of an high percentage of resident coming from non UE countries, mainly masculine and belonging to 30 – 50 age class. The tendency of Italian population ageing is confirmed. This process is softened, in the Porta Palazzo area, by the presence of non UE resident (see Figure 4).

FIGURE 3. – *Lingotto (left) and Porta Palazzo(right) areas.*

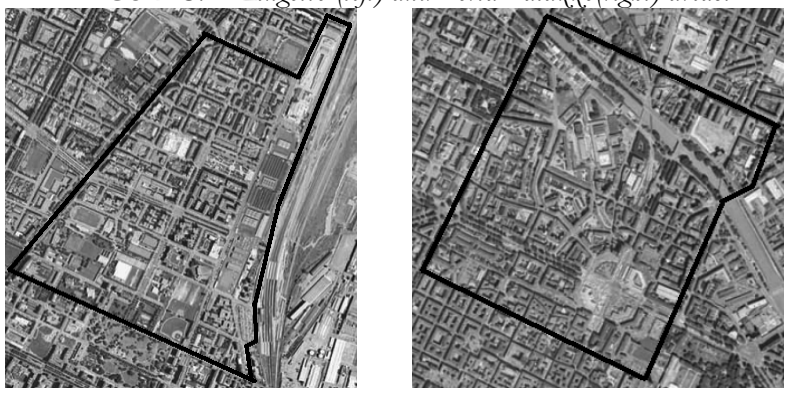


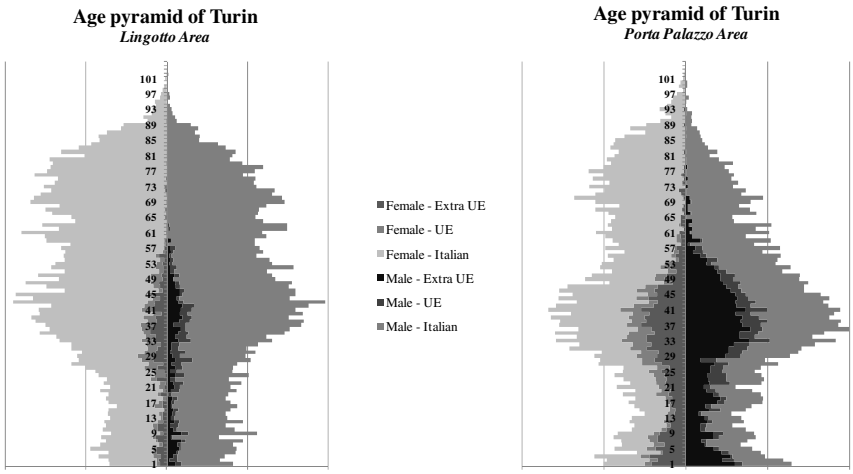
TABLE 2 – Demographic composition of Porta Palazzo and Lingotto areas. Population classified by geographic origin.

Origin	Lingotto	Porta Palazzo	Total
Italian	18.341	11.926	30.267
European (UE)	1.107	4.213	5.320
Extra UE	911	1.576	2.487
Total	20.359	17.715	38.074

TABLE 3 – Demographic composition of Porta Palazzo and Lingotto areas. Population classified by age class.

Age class	Lingotto	Porta Palazzo	Total
0 – 14	2.223	2.249	4.472
15 – 29	2.556	2.548	5.104
30 – 44	4.422	4.768	9.190
45 – 59	4.111	3.453	7.564
60 – 75	4.092	2.670	6.762
> 75	2.955	2.027	4.982
Total	20.359	17.715	38.074

FIGURE 4. – *Age Pyramid of Lingotto and Porta Palazzo areas.*



5.1 Multivariate analysis of the Space Syntax output

As stated earlier, it is quite rare to be able to give an aprioristic specification of the best factorial rotation. In this work some of the most used and effective rotations had been used (Varimax, Quartimax, Biquartimax and Equamax all in the raw and normalized versions). The rotations which gave the best results in terms of interpretation of the factor were Varimax and Biquartimax, both

normalized. The resulting factor loadings for the normalized Varimax rotation are shown in Table 2.

It may be relevant to outline that these computations had been done separately on all the axes of interest (the 130 axes of Lingotto and 148 of Porta Palazzo) and therefore no inferential procedure is required at this step as this is a census study. Moreover, as the Space Syntax measurements are done on the global city map the analysis has been performed on a unique dataset.

Although a number of variables are well identified with one single factor, some are slightly correlated with one or both the other two factors (Integration, Mean Depth and Node Count R2). The three factors identified represent approximately the 91% of the overall variability. The two solutions are quite similar and analyzing the loadings (Table 2), the interpretations given to the Factors are the following:

Factor 1 – (*Permeability*) Identifies long, well connected and strategically important axes that roughly constitute the radial structure of paths connecting the centre with the periphery of the system. Choice contributes significantly to the factor, making it a measure of what in graph theory is called betweennes. Because of the relevance of Connectivity and Control, the factor gets higher in axes with many connections and with more connections than their immediate neighbours. These elements together mean that the factor also represents the degree of permeability allowed by the axes to crossing paths. The factor is relatively stable from one case study to the other, with only Choice R2 remarkably changing its factor loading.

Factor 2 – (*Hierarchical subordination*) It inversely relates Mean Depth R2 and Controllability, which is not surprising since the former is a component in the function to calculate the latter. Because of that, the factor can be safely assumed to represent little else than controllability itself, which in itself shows little or no rela-

tion with the rest of Space Syntax Analysis measures. This factor is consistent in both case studies.

Factor 3 – (*Centrality*) This factor changes from one case study to the other. In the Porta Palazzo case study, the connection between Integration and Mean Depth is again somewhat expected since one is calculated through the other. Remarkably, however, Integration do not appear significant in factor 1 together with Choice, despite the fact that both Integration and Choice are Centrality measures. We can conclude that Integration (which in general graph theory would be considered a way to represent closeness) and Choice are different and non-redundant ways to represent Centrality. In the Lingotto case study, the behaviour of Choice R2 diverges from that of Choice, with the former being the main determinant of Factor 3. What was Factor 3 in Porta Palazzo case study is pretty much half way between Factor 1 and Factor 3.

TABLE 2. – *Factor Loading for the Varimax Normalized rotation. Shaded loadings whose absolute value is bigger than 0.65*

Quarter	**Porta Palazzo**			**Lingotto**		
Factor	**1**	**2**	**3**	**1**	**2**	**3**
Choice	0,90	0,03	0,26	0,92	0,13	0,17
Choice R2	0,92	-0,05	-0,01	0,18	-0,02	0,95
Connectivity	0,86	0,07	0,48	0,96	0,14	0,21
Control	0,89	0,12	0,38	0,96	0,16	0,09
Controllability	0,08	0,98	-0,18	0,19	0,96	-0,06
Integration	0,43	-0,41	0,79	0,62	-0,42	0,59
Line Length	0,87	0,04	0,45	0,93	0,11	0,30
Mean Depth	-0,32	0,46	-0,80	-0,57	0,50	-0,54
Mean Depth R2	-0,05	-0,97	0,21	-0,12	-0,98	0,09
Node Count R2	0,72	-0,22	0,63	0,75	-0,25	0,57
Expl.Var	4,77	2,35	2,36	4,91	2,43	2,04
Prp.Totl	0,48	0,24	0,24	0,49	0,24	0,20

5.2 Clustering by axes using the k-means algorithm.

The clustering of axes into groups is done using the approach of classification of subjects by employing the usual *k*-mean clustering technique to the composite variables obtained through the rotation in factor analysis (Takeuchi et al. 1982). The procedure follows a simple and easy way to classify a given data set through a pre-specified number of clusters *k*, therefore the problem of determining "the right number of clusters is generally a difficult task of considerable interest (Ming-Tso Chiang and Mirkin, 2007). Although there is not a unique solution to this problem, most of the authors (CIT) suggest an "heuristic" approach as the one used for the best factor rotation principle. The best results in classification and attribution of meaning to these clusters has been obtained using four groups. In Table 4 the group means and standard deviations for all the variables and factors analyzed are shown. It is quite evident the difference, in means, among the area, i.e. Lingotto and Porta Palazzo are urbanistically very different. A three dimensional representation of the axes is shown in Figure 5. As a matter of fact, this procedure clearly distinguishes the four main axes of the two areas: Via Milano, Corso Giulio Cesare, Corso Regina Margherita (in Porta Palazzo area), Corso Unione Sovietica and Corso Duca degli Abruzzi (in Lingotto area).

6. The Genoa case

The city of Genoa is an important seaport in northern Italy inhabited by about 610,000 people with an urban area population of about 900,000 units. Before 1100 Genoa emerged as an independent city-state becoming the most influent city in the Tyrrhenian Sea. Being the inland area of Genoa principally hilly, the urban structure developed over the centuries around the Old Harbour without an evident and precise structure. At the present day, the city of Genoa covers an area of approximately 243 square kilome-

TABLE 4. – Group means for all the variables (Turin Case).

LINGOTTO	1		2		3		4	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Choice	434,1	1.085,3	11.576,8	22.650,2	1.216,1	2.524,0	191.478,0	209.299,5
Choice R2	0,9	1,2	12,4	7,8	3,4	4,0	48,8	49,9
Connectivity	3,4	2,1	7,5	3,3	3,5	1,9	32,4	21,7
Control	0,8	0,4	1,5	1,0	0,6	0,4	7,2	6,1
Controllability	0,2	0,0	0,1	0,0	0,1	0,0	0,2	0,1
Integration	1,4	0,1	1,7	0,1	1,5	0,1	1,9	0,2
Line Length	255,5	270,3	739,0	498,1	256,8	178,9	2.865,0	1.849,7
Mean Depth	7,8	0,7	6,3	0,3	7,3	0,2	6,0	0,5
Mean Depth R2	1,7	0,0	1,9	0,0	1,9	0,0	1,8	0,1
Node Count R2	13,7	7,3	86,3	32,0	28,9	12,8	163,2	81,0
Valid N	10		16		27		5	
FACTOR 1	-0,5	0,2	0,2	0,3	-0,3	0,1	2,0	2,8
FACTOR 2	1,6	0,6	-0,9	0,5	-0,2	0,4	0,8	0,4
FACTOR 3	-0,2	0,2	0,3	0,4	-0,4	0,3	1,6	2,8

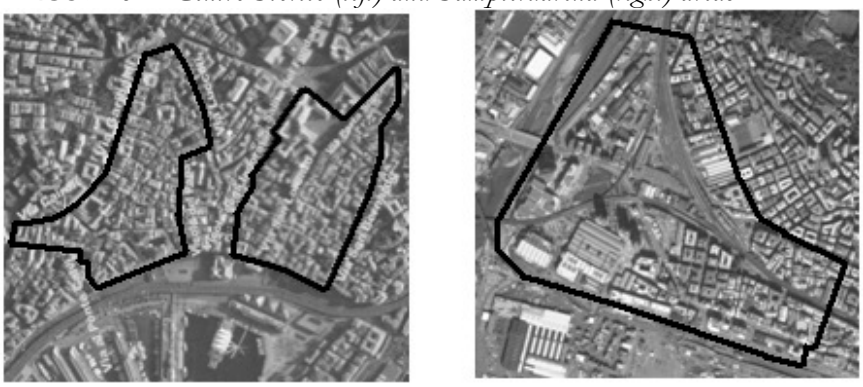
PORTA PALAZZO	1		2		3		4	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Choice	802,6	1.520,9	24.281,8	50.460,6	876,0	1.242,8	1.245.201,0	145.788,4
Choice R2	4,3	5,1	24,7	51,8	2,7	3,3	801,0	842,9
Connectivity	4,3	2,4	10,1	9,6	3,4	1,9	71,5	2,1
Control	1,0	0,5	1,8	2,1	0,7	0,4	14,4	2,9
Controllability	0,3	0,1	0,1	0,0	0,1	0,0	0,2	0,0
Integration	1,7	0,1	2,0	0,1	1,7	0,1	2,6	0,0
Line Length	152,0	146,3	625,2	661,1	160,7	120,8	5.992,1	508,3
Mean Depth	6,5	0,3	5,6	0,3	6,3	0,4	4,6	0,1
Mean Depth R2	1,7	0,1	1,9	0,0	1,9	0,1	1,8	0,0
Node Count R2	16,1	9,6	101,6	52,9	30,8	17,7	421,0	46,7
Valid N	19		38		30		2	
FACTOR 1	-0,2	0,2	-0,2	0,5	-0,1	0,2	8,3	2,1
FACTOR 2	1,5	0,7	-0,5	0,6	-0,3	0,6	0,4	0,2
FACTOR 3	-0,3	0,5	0,7	0,7	-0,8	0,6	-0,3	1,5

tres (151 sq miles) between the Ligurian Sea and the Apennine Mountains. The city develops on the coast for about 30 kilometres (18 miles) from east to west and for 10 kilometres (6 miles) from the coast to the north along the valleys Polcevera and Bisagno.

Just like in the Turin case, the areas of interest Maddalena (MD), S. Lorenzo (SL) and Sampierdarena (SP) are very different (Figure 6). Maddalena and S. Lorenzo are the two main areas of the Old Historical Centre and represent a big part of the medieval town made of narrow and winding allies (the so-called *vicoli*). Maddalena is one of the quarters which were constituting the original historical centre in Genoa and it is basically extended around the homonymous street. A bundle of narrow and streets compose this area

which can be assumed to be a stand-alone medieval suburb inside the city of Genoa. S. Lorenzo is another quarter in the Historic Centre of Genoa. Its structure has been deeply influenced over the centuries by the close presence of the port and many of the narrow streets which compose it are directed towards the sea. In the centre of this area lies a wide (in comparison with the contiguous streets) pedestrian road going from the Old Port to the old site of the Doge (the "Sire" of Genoa) passing on the right side of the S. Lorenzo Cathedral. Today S. Lorenzo is very close to the modern Genoa downtown.

FIGURE 6. – Centro Storico (left) and Sampierdarena (right) areas.



Sampierdarena was a fishermen village which has been annexed to Genoa in 1926. In the end of the 19th century, the availability of water from the Polcevera torrent supported the establishment of many factories in the area and Sampierdarena became one of the most important industrial areas in Italy. The industrial vocation of the area, although still visible, in now less relevant and the coexistence of a part of the harbour with both the highway and railroad junctions supported the development of intermodal transportation infrastructures.

Maddalena is inhabited by 3.896 residents, S. Lorenzo by 4.457 residents and Sampierdarena by 10.849. The demographic composition of the three populations, as we can see in Table 4, Table 5 and Figure 5, presents different distributions. The Maddalena area is characterised by an higher percentage of non-Italian residents corresponding to about 21% of the total population (S. Lorenzo presents 17% and Sampierdarena 18,6% of non-Italian residents). The age distribution of the three areas of interest is very different. The Maddalena area is characterised by an higher percentage of young people, only 22% of the residents are sixty or more. This percentage increase in S. Lorenzo area (26%) and more over in Sampierdarena area (33%).

The age pyramids of the three areas highlight the previous considerations. The different age percentage distribution is well rendered. The age percentage distribution of the Sampierdarena area is clearly related to an ageing population. This is not the same for the two quarters of the historical center of Genoa, Maddalena and S. Lorenzo. In fact, observing the two population pyramid (Figure 5), we can see two younger populations.

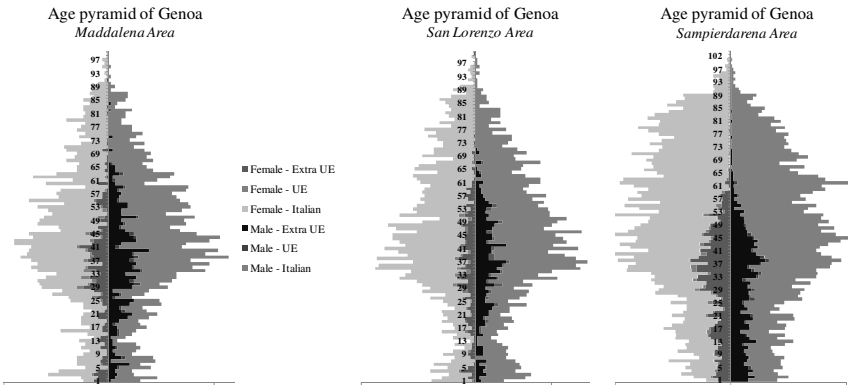
TABLE 4 – Demographic composition of Maddalena, S. Lorenzo and Sampierdarena areas. Population classified by geographic origin.

Origin	Maddalena	S. Lorenzo	Sampierdarena	Total
Italian	3.083	3.700	8.835	15.618
European (UE)	720	658	1.857	3.235
Extra UE	93	99	157	349
Total	3.896	4.457	10.849	19.202

TABLE 5 – Demographic composition of Maddalena, S. Lorenzo and Sampierdarena areas. Population classified by age class.

Age class	Maddalena	S. Lorenzo	Sampierdarena	Total
0 – 14	403	459	1.119	1.981
15 – 29	514	539	1.444	2.497
30 – 44	1.165	1.270	2.377	4.812
45 – 59	944	1.052	2.349	4.345
60 – 75	562	722	2.060	3.344
> 75	308	415	1.500	2.223
Total	3.896	4.457	10.849	19.202

FIGURE 5. – *Age Pyramid of Maddalena, San Lorenzo and Sampierdarena areas.*



6.1 Multivariate analysis of the Space Syntax output.

The multivariate analysis of the 290 axes (Fiumara: 95; Maddalena: 104; S. Lorenzo: 91) gave results, summarized in Table 5, which can be considered comparable with the ones computed for the Turin case (Table 2) although some differences are present. The general meaning of the factors for the Genoa case can be assumed to be very similar to the Turin case. Sampierdarena is the

quarter which for urban layout is the most similar to Porta Palazzo or Lingotto and the fact that the composition of the three factor is exactly the same of Turin seems to be a validation of the meanings given to the loading shown in Table 2. It is is interesting to see how the choice variable is not firmly constituting any factor for the two area in the "Centro Storico" of Genoa.

Factor 1 – (*Permeability*) Given the number of variables recurring from the Factor 1 identified in the Torino case studies, we can argue that it mainly conveys the same meaning. However, while in the Sampierdarena case study we have an orthogonal grid, much like those found in Torino, and the Factor is correspondently very similar, the other case studies of Genova have an evidentely different kind of urban grid and the Factor itself looks to be influenced by this. Both in the Maddalena case study and, less evidently, in the San Lorenzo case study, Choice is less connected with factor 1 and more closely related to Integration.

Factor 2 – (*Hierarchical subordination*) There is no significant difference in this factor from what we found in the Torino case studies.

Factor 3 – (*Centrality*) Except for the different behaviour of Choice, it closely resembles the Factor 3 we found in the Porta Palazzo case study and mainly consists of Integration and its component Mean Depth.

6.2 Clustering by axes using the k-means algorithm.

Using the three identified factors for clustering the axes into four categories using the *k*-means algorithm produced the results shown in Table 6. The labelling 1-4 for the groups has been done in order to identify the corresponding group category with the Turin case. Generally it can be stated that the group means for each factor and for each variable show a similar pattern to that we found in To-

rino. A cluster includes the main roads of each area, where Factor 1 and Factor 3 are both much above the average. A second cluster is at the other end of the spectrum, with the lowest or next to lowest averages for Factor 1 and Factor 3. The other two clusters show subtler differences, with a high average of Factor 1 and a low average of Factor 3 or vice versa.

TABLE 5. – *Factor Loading for the Varimax Normalized rotation for the three areas of interest in Genoa. Loadings whose absolute value is bigger than 0.65 are shaded.*

Quarter	Maddalena			S. Lorenzo			Sampierdarena		
Factor	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Choice	0,13	0,24	0,71	0,40	0,04	0,50	0,72	-0,15	0,08
Choice R2	0,90	0,11	0,03	0,84	0,12	0,22	0,84	0,21	0,14
Connectivity	0,82	0,28	0,45	0,94	0,08	0,26	0,94	0,19	0,19
Control	0,82	0,42	0,30	0,88	0,36	0,12	0,90	0,35	0,09
Controllability	0,20	0,97	-0,05	0,16	0,97	-0,14	0,16	0,97	-0,11
Integration	0,23	-0,21	0,89	0,26	-0,23	0,92	0,20	-0,14	0,97
Line Length	0,66	0,05	0,64	0,87	0,00	0,33	0,89	0,19	0,11
Mean Depth	-0,23	0,22	-0,88	-0,25	0,22	-0,92	-0,18	0,14	-0,97
Mean Depth R2	-0,12	-0,97	0,09	-0,07	-0,98	0,17	-0,08	-0,97	0,17
Node Count R2	0,68	-0,34	0,56	0,77	-0,42	0,39	0,84	-0,39	0,25
Expl.Var	3,25	2,42	3,11	4,04	2,32	2,39	4,53	2,34	2,07
Prp.Totl	0,33	0,24	0,31	0,40	0,23	0,24	0,45	0,23	0,21

7. Comparing the multivariate analyses for the two cases

With the differences between the factor loadings in Torino and Genova in mind, we can attempt an interpretation of what was found.

First of all, the interrelation between two different measures of centrality like Choice and Integration is a pretty complex one. While at times they have a direct proportionality of some sorts, with both being high in very important axes and both very low in a few marginal ones, other times they can also produce contradicting indications. This should be probably seen as an indication that the two measures are not redundant and, while trying to describe centrality, they end up showing properties of space and configuration that are at least partially different.

Choice, in particular, and its ambivalent relation with Factor 1, should be regarded as a very significant insight on the mechanics of Space Syntax.

Choice is a global measure that, in fact, depends on local as well as global properties of the grid.

It measures the number of times a given axis stands on the shortest path between the couples of every other axis in the grid

In an axial map with a large enough number of axes, most of them will only represent a shortest path between the few departures and destinations that stand very close to the axis itself. The likelihood of long, trans-urban paths actually going through each axis will, on the other hand, be pretty low.

This means that for most of the axes in a large axial map, a big part of their Choice value (if not the entirety) will come from couples that are either directly connected to it or stand at 1, 2 syntactic steps of distance.

In other words, the higher the connectivity, the higher the chance that this local component of Choice will be above zero.

However, a high connectivity is not*, per se*, a sign of proximity to the centre of the system, and consequently there is no guarantee that a value of Choice that is built on short, local paths, will be in any way associated with a proportional value of Integration.

TABLE 6. – Group means for all the variables.

FIUMARA	1		2		3		4	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Choice	96.895,2	309.231,3	466.221,7	1.340.761,5	170.304,5	634.913,6	4.888.092,3	3.695.449,9
Choice R2	3,7	6,8	5,6	8,6	1,8	2,3	83,3	58,6
Connectivity	3,8	2,3	4,0	2,2	2,7	1,2	21,3	5,6
Control	1,3	0,8	0,8	0,7	0,7	0,4	7,2	2,8
Controllability	0,3	0,1	0,2	0,1	0,2	0,1	0,3	0,1
Integration	0,4	0,0	0,4	0,0	0,4	0,0	0,4	0,0
Line Length	161,5	242,4	144,1	108,4	100,3	55,5	998,3	240,2
Mean Depth	26,0	1,0	24,9	0,6	27,6	0,7	24,7	1,5
Mean Depth R2	1,6	0,1	1,8	0,1	1,8	0,1	1,7	0,1
Node Count R2	11,1	5,7	27,1	13,6	17,5	12,0	66,7	8,4
	26		44		19		6	
FACTOR1	-0,4	0,4	-0,2	0,5	-0,1	0,3	3,4	0,8
FACTOR2	1,1	0,5	-0,6	0,5	-0,4	0,9	0,5	1,2
FACTOR3	0,0	0,7	0,6	0,4	-1,5	0,6	0,2	1,3

MADDALENA	1		2		3		4	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Choice	29.676,8	103.858,2	5.343,8	8.629,1	189.534,3	641.018,1	5.066.149,8	6.326.783,1
Choice R2	7,3	10,1	3,8	5,7	6,3	9,1	71,2	86,0
Connectivity	4,8	2,5	3,4	1,6	4,7	1,9	16,0	4,4
Control	1,3	0,6	0,7	0,4	0,9	0,5	3,8	1,3
Controllability	0,3	0,0	0,2	0,0	0,2	0,0	0,3	0,0
Integration	0,5	0,0	0,5	0,0	0,5	0,0	0,5	0,0
Line Length	62,9	33,0	50,7	18,9	104,8	45,9	278,0	90,8
Mean Depth	22,7	0,7	23,0	0,4	21,7	0,4	21,1	0,7
Mean Depth R2	1,7	0,1	1,8	0,0	1,8	0,0	1,7	0,0
Node Count R2	15,6	7,7	17,9	7,2	31,2	8,8	51,8	9,8
	28		36		35		5	
FACTOR1	-0,3	0,6	-0,6	0,4	0,6	0,6	1,9	2,7
FACTOR2	1,2	0,5	-0,2	0,5	-0,9	0,6	1,1	0,9
FACTOR3	-0,2	0,7	-0,2	0,5	0,0	0,5	2,6	3,0

S. LORENZO	1		2		3		4	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Choice	15.903,2	25.843,5	127.505,8	558.933,2	16.406,9	23.116,3	3.857.649,0	5.119.834,8
Choice R2	4,6	4,7	4,1	4,1	4,5	5,9	35,8	12,6
Connectivity	3,9	1,7	4,1	1,7	4,3	2,5	12,8	3,5
Control	1,0	0,4	0,7	0,3	0,9	0,6	2,6	0,9
Controllability	0,3	0,0	0,1	0,0	0,2	0,0	0,2	0,1
Integration	0,5	0,0	0,5	0,0	0,5	0,0	0,5	0,0
Line Length	62,1	38,2	76,7	39,7	67,2	38,0	246,9	99,6
Mean Depth	22,6	0,6	21,8	0,5	23,5	0,4	21,2	0,7
Mean Depth R2	1,7	0,1	1,8	0,0	1,8	0,1	1,8	0,1
Node Count R2	14,2	5,9	28,7	12,7	22,5	9,9	55,0	11,0
	33		35		15		8	
FACTOR1	-0,4	0,5	-0,3	0,5	0,3	0,7	2,5	1,0
FACTOR2	1,0	0,6	-0,8	0,5	-0,5	0,6	0,2	0,8
FACTOR3	0,0	0,6	0,5	0,6	-1,5	0,3	0,9	1,5

Such uncertainty accounts for Integration having a factor loading on Factor 1 in Torino between 0.4 and 0.6.

In a regular, almost orthogonal grid, Choice is mainly a product of Connectivity (except for a few, very important axes) and may or may not be related to Integration.

A totally different situation comes when the grid is less orthogonal, more fragmented and dispersed.

In such a case, short axes with few connections and marginal levels of local Choice, can actually be the sole connection (or one of the few) between different parts of the graph.

When that happens, Choice loses its relation with Connectivity and extremely increases that with Integration.

Unsurprisingly, this is what happens in the two case studies from the historical city centre of Genoa, which fits the image of fragmented urban grid pretty well.

8. Conclusions and further work

Most of the international analyses and surveys on crime are not taking into account the key role of the urban layout. Space Syntax offers a quantitative and objective way to analyze the structure of a city measuring a set of variables whose usage for the definition of the sampling designs seems to be quite straight.

The results obtained in the Turin case show that the factor analysis for the space syntax variables gives clearly interpretable composite variables. Moreover, the *k*-means clustering technique distinguishes four groups of axes structurally different. Although this work is only a part of the research project described in paragraph 1 it analyses one of its key points. Many are the possible evolutions of the present work:

- analysis of the space syntax variables for other areas and comparison of the factor loadings and meanings;
- usage of other multivariate approaches for the composite variables definitions (e.g. Multidimensional Scaling) and comparison of the corresponding results with the ones given by the factor analysis;
- evaluation of Space Syntax as a predictor of pedestrian flow (amount and composition) under different conditions and through different indicators;
- measurement of other urban (e.g. height of buildings, number of buildings, number of building main doors, number of abandoned buildings), social (e.g. nationality and age composition per axis) and economic variables (number of shops, number of shop windows per axis, number of security cameras) in order to evaluate if the axis clustering procedure based on the space syntax measurements is effective also for other phenomena;
- study of the perception of safety for people living, working (or both) in different axis groups: how the different structure of the axes may influence this perception;
- comparison of the perception of safety between people leaving or working in the areas of interest and pedestrians who are simply crossing axes.

All these topics require a strict definition of the urban axis and a clustering principle which have been herein discussed.

REFERENCES

Arabie, P. (1977). Clustering representations of group overlap. *Journal of Mathematical Sociology*, 5:113-128.

Bock, R.D., and Husain, S.Z. (1952). Factors of the tele: a preliminary report. *Sociometry*, 15:206-219.

Burt, R.S. (1984). Network items and the general social survey. *Social Networks,* 6:293-340.

Caldeira, G.A. (1988). Legal Precedent: Structures of Communication Between State Supreme Courts. *Social Networks*, 10:29-55.

Cattell R.B. (1978). *The Scientific Use of Factor Analysis in Behavioural and Life Sciences*, Plenum Press.

Cicchitelli, G., Herzel, A., Montinari, G.E. (1992). *Il Campionamento Statistico*, Il Mulino, Bologna.

Collins, R. (1988). *Theoretical sociology*. Harcourt Brace Jovanovich, New York.

Diana, G. and Salvan, A. (1993*). Campionamento da popolazioni finite*. Cleup, Padova.

Durkheim, É. (1897). Suicide. The Free Press (Reprint 1966), New York.

Erikson, B. (1988). *The relational basis of attitudes*. In Willman, B. and Berkowitz, S.D. (eds.), *Social Structures: A Network Approach*, pages 99-121. Cambridge University Press, Cambridge.

Everitt, B. (1974). *Cluster analysis.* Heineman Educational

Fabbris, L. (1989). *L'Indagine Campionaria.* La Nuova Italia Scientifica, Roma

Freeman, L.C., Romney, A.K., and Freeman, S. (1987). Cognitive structure and informant accuracy. *American Anthropologist,* 89:310-325

Friedkin, N.E. (1984). Structural Cohesion and Equivalence Explanations of Social Homogeneity. *Sociological Methods and Research*, 12:235-261

Gambini, A. (2009). *Il Campionamento Statistico per la Ricerca Sociale e di Mercato*. Giappichelli.

Hillier, B. and Hanson, J. (1984). The Social Logic of Space. Cambridge University Press.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

Joliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag.

Kalton, G. (1989). *Introduction to survey sampling*. Sage Publications.

Kruger, M.J.T. (1989). *On node and axial grid maps: distance measures and related topics*. Unit for Architectural Studies, Bartlett School of Architecture and Planning, University College of London.

Laumann, E., Pappi, F. (1973). New Directions in the Study of Community Elites. *American Sociological Review*, 38: 212-30.

MacRae, D. (1960). Direct factor analysis of sociometric data. *Sociometry*, 23:360- 370

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press

Ming-Tso Chiang, M. and Mirkin, B. (2007). *Experiments for the Number of Clusters in K-Means,in Progress in Artificial Intelligence*, 13th Portuguese Conference on Artificial Intelligence, EPIA 2007.

Press, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston Inc.

Ratti C. (2004). Space syntax: some inconsistencies. *Environment and Planning B: Planning and Design* **31**:487–499

Takeuchi, K., Yanai, H. and Mukharjee, B.N. (1984). *The Foundations of Multivariate Analisys*. Wiley Eastern Limited. New Delhi.

Teklenbur, J.A.F., Timmermans, H.J.P. and van Wagenberg A.F. (1993). Space syntax: standardised integration measures and some simulations. *Environment and Planning B: Planning and Design*, 20:347-357.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Application,*. Cambridge University Press, Cambridge.

Weber, M. (1958). *The City*. The Free Press, New York.

Wille, R. (1984). *Line diagrams of hierarchical concept systems*. Int. Classif. 11, 77-86

Wright, B. and Evitts, M.S. (1961). Direct Factor Analysis in Sociometry. *Sociometry,* 24:82–98.

Working Papers recently published

(The complete list of working papers can be found at http.//www.disefin.unige.it)

n.4/2011 E. Briata, "Marginal tax rates, tax revenues and inequality. Reagan's fiscal policy", July 2011

n.3/2011 Francesco Copello, Cristiana Pellicanò, "Esemplificazione della Data Envelopment Analysis per la valutazione di efficienza in una grande azienda ospedaliera universitaria"

n.2/2011 Stefano Capri, Rosella Levaggi, "Shifting the risk in pricing and reimbursement schemes? A model of risk-sharing agreements for innovative drugs"

n.1/2011 Cinzia Di Novi, "The Indirect Effect of Fine Particulate Matter on Health through Individuals' Life-style"

n.4/2010 Angelo Baglioni, Andrea Monticini, "Why does the Interest Rate Decline Over the Day? Evidence from the Liquidity Crisis"

n.3/2010 Amedeo Fossati: "The double taxation of savings: the Italian debate revisited"

n.2/2010 Andrea Monticini, David Peel, Giacomo Vaciago: "The impact of ECB and FED announcements on the Euro Interest Rates"

n.1/2010 Amedeo Fossati: "Vilfredo Pareto and the methodology of the Italian tradition in public finance"