Master thesis

A novel GIS method to determine an urban centrality index applied to the Barcelona metropolitan area

Timothée PRODUIT
June 2009

Professors:
F. GOLAY, S. PORTA
Supervisors:
N. LACHANCE-BERNARD, S. JOOST, E. STRANO

EPFL - SSIE - LaSIG

PoliMi - HSL

Aknowledgements:

*Ce travail de diplôme a été réalisé à l'*EPFL (CH) *en collaboration avec le* PoliMi (IT)*.*
Ainsi mes remerciements vont au laboratoire LaSIG et à ses collaborateurs parmi lesquels N. Lachance-Bernard, S. Joost, F. Golay, J. Ingensand, M. Kalbermatten, E. Andrey *et* V. Boillat Kireev *n'ont jamais compté leur temps pour contribuer à la réalisation de ce travail.*
*Escaladant les Alpes pour se laisser glisser vers la plaine de Milan, je tiens à remercier le laboratoire HSL (*S. Porta, E. Strano *et* E. Morello*) qui m'accueilli et m'a aidé à trouver les bonnes directions pour mener à bien ce diplôme.*
Si le climat du sud était bien agréable ce printemps, il me faut faire le chemin inverse, retrouver les montagnes pour remercier encore ma famille qui m'a soutenu durant les cinq années qui ont mené au diplôme.

Contents

Abstract

Events occurring in a city, such as car accidents, attacks on the security or activities locations are channelled in the urban network. The goal of this report is to provide a GIS tool able to compute densities and diversities in a network space rather than in the euclidean one.

The network space created to support these calculations is a *shortest path tree* for a given bandwith. On the basis of this shortest path tree, three parameters are calculated: the density and the diversity of activities and the density of edges. The case study of this report is the city of Barcelona. The activities of this city were projected on the network. Indeed, activities like other events in the urban space tend to follow the network.
The *Human Space Lab* (HSL) of the *Politecnico di Milano* (IT) has applied its *Multiple Centrality Assessment* to Barcelona's network. Thus, edges become a value of centrality. The network density of the edges presented in this report will take into account these attributes such as the value of population used in other available GIS density tools. The further idea, is to weigh the events (activities or edges) by their distances, as in a standard kernel density estimation.

The network approach reveals a spatial distribution of the values sensitive to the favourite direction of the network. The diversity measured along the network make out particular area of the city as residential or commercial zones.

Résumé

Les événements qui ont lieu dans une ville qu'ils soient des accidents de la circulation, des atteintes à la sécurité ou simplement des localisations d'activités sont canalisés dans le réseau urbain. Le but de ce rapport est de fournir un outil SIG capable de calculer des densités et diversités dans un espace réseau plutôt que dans l'espace euclidien.

L'espace réseau créé comme support à ces évaluations est un *arbre des plus courts chemins* d'une longueur précisée. Sur la base de cet arbre, trois paramètres sont calculés: une densité et diversité d'activités et une densité de segments. L'étude de cas de ce rapport est la ville de Barcelone. Ainsi, les activités de cette villes ont été projetées sur le réseau. En effet, les activités comme les autres événements de l'espace urbains ont tendance à suivre le réseau.
L'*Human Space Lab* (HSL) du *Politecnico di Milano* (IT) a appliqué son processus *Multiple Centrality Assessment* au réseau de Barcelone. Ainsi, chaque segment de la ville a reçu des valeurs de centralité. L'estimation de densité sur le réseau, présentée dans ce rapport, prend en compte ces valeurs de centralité comme une valeur de population que l'on retrouve dans d'autre outils SIG d'estimation de densité. Cette estimation de densité se fait encore en pondérant les événements par leurs distances comme cela est fait avec une estimation de densité de kernel standard.

L'approche réseau des indices amène une distribution spatiale des valeurs sensible aux directions privilégiées du réseau. La diversité mesurée sur le réseau permet de démarquer des aires particulières dans la ville (zones résidentielles, zones commerciales).

Introduction

Networks appear all around us and in most of our common daily tasks. Neuronal networks allow us to interact with our environment, electrical networks supply energy to our computer, which is linked to the *World Wide Web*. Without *Internet* how would we increase and represent our social network? For the ones which are courageous enough to leave their sofa and living-room, they will enter the urban network.
All these networks have been studied in the different fields they involve. Common characteristics with indexes and their statistical distributions have been found. Urban network have also been studied, mainly by urbanists.

The research laboratory *Human Space Lab* (HSL) from the *Politecnico di Milano* (IT) has created a tool which is able to charaterize nodes and edges of a network with centrality indexes in a primal way. However, centralities have to be compared with urban reality. With this goal, a kernel density estimation (KDE) has been applied to the location of the Barcelona (SP) and Bologna (IT) activities. This same KDE has been applied to the edges of this city weigthed by their centrality index. This approach allows to create a continuous field of values for the city area (Fig 1, page ix).
Results are already showing a high correlation between activity densities and some of the centrality indexes. Nevertheless, the kernel density estimation uses the euclidean space, but this space does not fit in the urban reality. In a city, the shortest path between a person's flat and the nearest fastfood is rarely a straight line. Urban trips are channeled in the urban network, as ponctual events like car accidents or commercial locations are.
The first goal of this report is to provide a *tree of the locations accessible with a fixed distance*. In other words: "Which location can I reach by walking 200m, and which edges can I use?". Following the Tobler first law of geography: *"Everything is related to everything else, but closer objects are more related than distant ones."* [38]

a distance weighted density is calculated. The diversity or entropy index will be computed for the activities belonging this shortest path tree.

The report, first, will present the objectives and method applied to create these indexes. Then, it goes through the search studies following the same or shared goals with this one. Next, the necessary theoretical basis are provided. Thus, the reader is able to understand how the tool have been developed. After that, two case studies are presented. The first one, on a local scale explains how to understand results of the network calculations. The second one, on a global scale, give a visual analysis between the different parameters calculated in Barcelona. Finally, analysis and review of the indexes conlude the report.

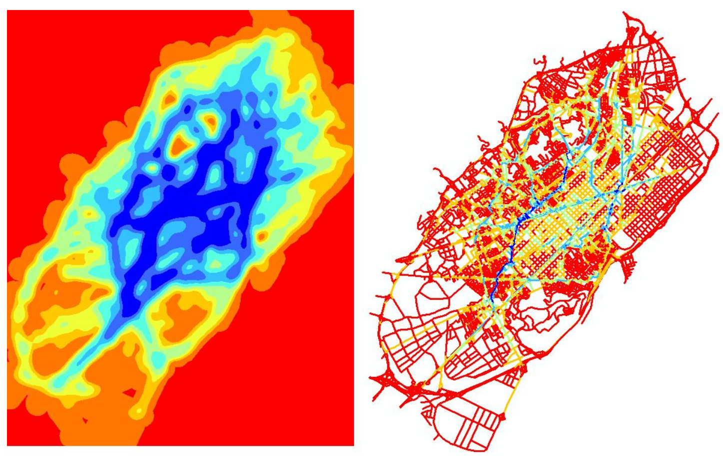


Figure 1: Right: Global betweenness centrality for Barcelona's network;
Left: Kernel density estimation of the global betweenness.

Chapter 1

Problematic, objectives and method

1.1 Problematic

The Multiple Centrality Assessment (MCA) is a procedure of network analysis which qualifies the edges centrality on an urban network. These indexes are well understand in topological network. Nevertheless, the urban centrality has to be understood in the large context of the city. Does them reflect how the city grow, architecturally and economically develop? Does the street centralities correlate with land use?

In order to correlates activities (points) and streets (polylines) with centrality attributes, a kernel density estimation generates density values in every raster cell of a map. This density estimation has two main characteristics. First, it weights the events (point or polyline) by their euclidean distances according to a bandwith. Then, it gives the possibility to take into account a "population" which can be a centrality index. Once the density of street centralities and activities are calculated, a correlation study allows to extract the relation between centralities and activities[1]. The density explains how many activities are counted in an area, but don't gives informations about the kind of activities. A value of diversity could explains how many different activities are located in this same area.
On the other hand, euclidean distances which are the basis of the standard GIS tools, don't fit with urban routes. Indeed, trips in a city are channeled in the urban network.

This problematic leads to the main goals of this report.

[1] A such study has been applied to Bologna (IT) [29].

1.2 Goals

In this way, the goals of this master thesis are multiple, but essentially are aimed at the implementation of new indexes:

- The first goal of this work is to give a new indicator for the land-use. This indicator will catch the diversity of activities.

- Then, a new distance, the network distance, closer to the urban routes than the euclidean distance is created.

- This network distance is the basis for the implementation of a *network density estimation* and a *network diversity* assessment.

- All these new parameters are applied to a real network: the city of Barcelona.

To reach these goals, several methods are choosen or developed.

1.3 Method

There is number of diversity indexes in many field. However, the diversity is a well known index especially in the biological field. The measure of entropy, initially developed for the information and communication systems and then widely applied to the biodiversity, is used here to measure the diversity of the activities.
To create network distances, the Dijkstra's algorithm allow to find the shortest path between to point of a network. A variation of this algorithm create a *shortest path tree* for a raster cell and a given bandwith. On this tree, the nodes have a value of distance which allows to find the network distance from a start point to an edge or a projected activities. These distances are the input for the *network density estimation* which is a *kernel density estimation* applied to network distance rather than euclidean one.
In the same, way projected activities belonging the shortest path tree are the input for the measure of the network diversity of activities.

The main algorithm is written within *Python language* and access to a *PostgreSQL/PostGIS* database to create spatial requests. The grid cells results are fed back to *ArcGIS* to map the values and generate the raster.

Preliminary note:

In this report the ***kernel density estimation*** (KDE) which is a *ArcGIS* tool, will be compared with the ***network density estimation*** (NDE) which is the tool developed in this work.
The *network density estimation* can be applied to a network, to avoid an ugly *network density of the network*, we will rather use *network density estimation* of edges. The *kernel density estimation* can be applied to networks as well, we will use then *kernel density estimation* of edges.

Thus, *network density* will refer to the tool NDE, and not to the *density of the network*...

Chapter 2

State of the art

This chapter is a review of scientific researches sharing with this report same or partial goals. First, the network centrality will be presented with a focus on urban centralities. Then, some works dealing with the kernel or network density are quoted and briefly explained. Diversity or entropy have been used in other tasks than biological studies, the final part will have a look on these researches.

2.1 Network and GIS

Network models are an important data model in GIS, they are involved in transportation, hydrology, communication or social connection with spatial component. All of them have a similar structure. The properties from the network becomes from the graph theory. Thus, topological properties of graph (connectivity and adjacency) are those that are not altered by elastic deformation. Difference can be done between tree, Manhattan (self-speaking) and hub-and-spoke networks where the edges radiate from a central vertex.

The complexity of the network models as implemented in GIS have evoluted. The first models or "spaghetti" do not preserve topological properties and is useless for network analysis. Once the topological properties of the network is established into GIS, routing can be applied. The linear referencing, allows to find the location of an event according to its location along an edge rather than with its coordinates. Most of the networks deal with routing problematic, thus they must support concepts of capacity and flow direction. Network modeling are still in move to answer the analytical needs of multi-modal networks, mobile communication, sensor networks... [22]

2.2 Network centrality

A purpose of this work is to generalize values of centralities*, which are attributes of the network segments, to the entire city field. So, the centrality is not the stake in this report but rather the pretext. The centrality of the urban networks is the object of numerous studies. The diverse directions taken by these studies are presented below. Then, a focus on the* ***Mulitple Centrality Assessment*** *(MCA) developed by* Porta *will be done. These MCA indexes will be the input for the* ***network density estimation****.*

The chemical and biological systems, the neuronal networks, the social interactions, the World Wide Web are so many examples of systems consisted by a large number of dynamic units highly interconnected [6]. The current studies aim at characterizing these networks, at finding parallels and at creating models which imitate their developments. The new available IT power allowed to set up a variety of different variables in this purpose.
Basic tools for these studies are the work of Bavelas who, in 1950, realizes that social relationships can be seen as social networks. To give a rigour to this hypothesis, its will is then to go from the metaphor to a formal and systematic study of these networks [4]. So, his study takes the direction of the graphs centrality. Later Freeman continues the researches of his predecessor to define a set of centrality indexes (*degree, betweeness, closeness*) [18].

It was demonstrated that all the networks quoted previously, possesses common characteristics. Among them, the said *Small-World* property for whom the average of the topological distances between nodes is weak with regard to the size of the network [35]. Furthermore, it was shown that most part of the real networks are *scale-free* as well, that is they are characterized by the presence of kernels which possess a number of connections (or *degree*) much bigger than the average of the other nodes (Fig 2.1, page 6). This statistical distribution of the nodes degree makes them strong against the errors and unpredictable attacks, because few nodes possess a central importance. The urban networks possess these characteristics as well and so, common tools can be applied to the urban system.

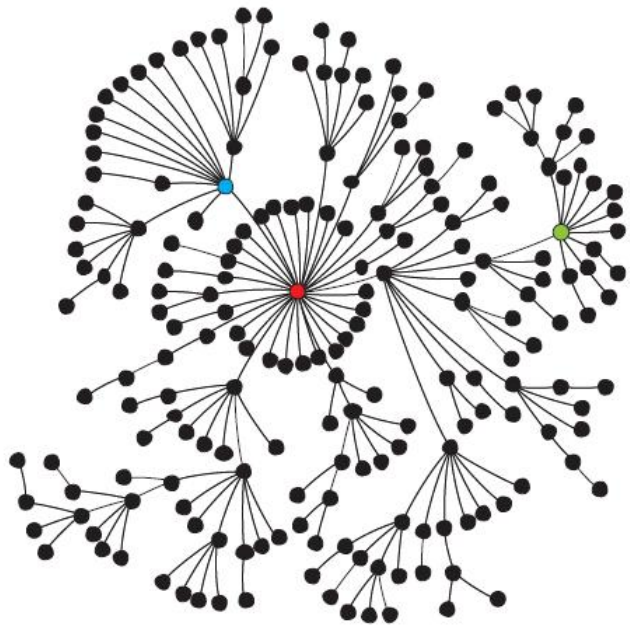


Figure 2.1: *Scale-free graph*, grown by attaching new nodes at random to previously existing nodes. The probability of attachment is proportional to the degree of the target node; thus richly connected nodes tend to get richer, leading to the formation of hubs. Colours indicate the three nodes with the most links (red, k = 433 links; blue, k = 412; green, k = 411). [35]

The network approach was widely used in the urban studies in particular to study the correlations between the land-use or traffic streams with transport networks. However, the application of the network approach to the urban public road system lead to number of question. How to take into account the distances? What sort of graph to use? What type of measures must be made? How interact the measures of structure and those of the dynamics of the network? What are the contributions that can make the GIS community [30]?

In the study of the centrality, it is necessary to differentiate two approaches. The first one, **primal** (*primal approach*), is based on the metrics of the geographical networks. The network of streets and intersections is then represented by a **spatial graph** in which the intersections are transformed into nodes and streets into segments in an euclidean space. It is the natural and intuitive way a network is represented in GIS. The second, **dual** (*dual approach*), is based on the study of **topological networks**, namely "*space syntax*" which is based on the assertion that certain locations possess a more important role because they are more *central* [27]. Streets are then represented by nodes and the intersections by segments in a connectivity graph (Fig 2.2).

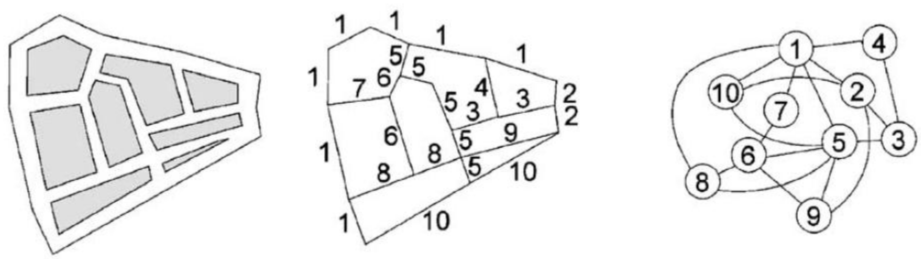


Figure 2.2: Generalization of the network of the city according to Jiang, street names are replaced by numbers. To the left the fictitive urban system, in the center the primal model of network , to the right the dual model of network [30]

2.2.1 Primal approach

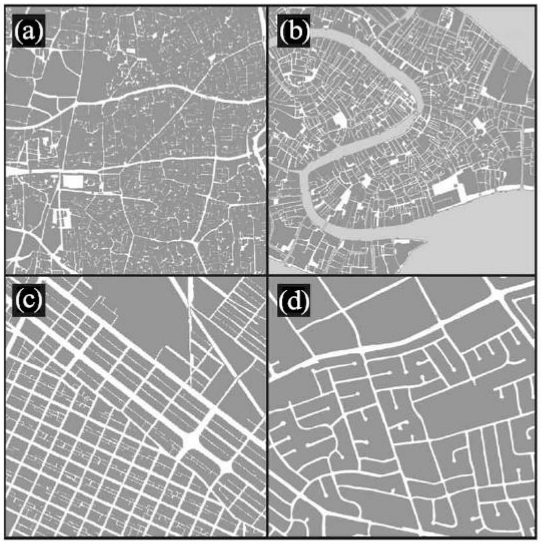
The primal or direct representation of an urban network is the intuitive way of representing the streets of a city in the space. It is this approach which was most widely applied to the networks of the territorial system. Indeed, this is the simplest method to integrate into the representation the important characteristic that is the distance. This intuitive representation is the one which corresponds best to the human perception (Fig 2.3(b), page 9). So, from the beginning the 60s, many researches were done with the aim of modelling the land-use, the behavior of the market, the streams of trips according to the topological and geometrical characteristics of the urban network. It allowed to find topological characteristics common for cities of similar constructions as well [12, 14]. Applied to the public analysis of the transport system of Australian cities, it was used to demonstrate the strengths and the weakness in the connectivity of these networks [31].

2.2.2 Dual approach

The dual approach is the one which was most widely used in the field of *urban design*. The *Space Syntax*, from the end of the 80s, after the publication of Hillier and Hanson, developed measures based on the graphs to analyze and understand

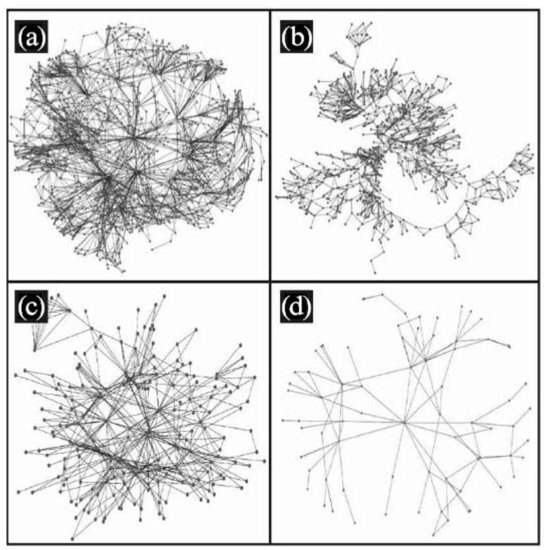
the complexity and the structure of the urban network. It established significant correlations between the topological accessibility of streets and phenomena as streams of trips, security against the crime, micro economic vitality, the separation of the activities, pollutions [20].
The main process which governs this approach is the conversion of the direct representation of the city in a connectivity graph. During this phase, the straight segments of street (*axial lines* or *line of sight*) are transformed into nodes and the intersections between axial lines in segments. The measures of accessibility are made on the basis of the connectivity graph and in a topological system of **non-euclidean** distances (FIG 2.3(c), page 9). The calculated values are then represented on the city map.

Applications of this approach covers number of disciplines as the modelling of the pedestrian trips, the streams of vehicles, mapping of the crimes... JIANG proposes the use of this approach for a model of networks generalization. The application of the measures of centrality which considers at once the global and local scales of streets grouped according to their names, allows to select the main segments in a purpose of mapping. He advances that this approach, able to extract the main structure of a network, is of interest for all the fields of application which base on this structure (built environment, electrical and gas network) but also the socioeconomic activities in the city. However, its generalization by street name was criticized for its introduction of nominal values in a spatial context [21].



(a) Original maps.

(b) Primal graph: Streets are edges, intersection are nodes.



(c) Dual graph: Streets are nodes, intersections are edges

Figure 2.3: Four samples of 1 square mile for the cities of Ahmedabad (a), Venise (b), Richmond (c) et Walnut Creek (d) [27].

2.2.3 One-on-one debate of the approaches

These two previous approaches were used in urban studies. Each has advantages and inconveniences. The main criticism made for the dual graphs is the process of aggregation. During this phase, which reduces streets to nodes, the network lose the information of distance and makes a wide place for subjectivity. However, the distance is important because it is it who allows to estimate the density of intersection for example. In the same way, it is not possible to study the variation of the number of intersection along the segment of the network.
Its advantage is mainly that once transformed into a dual graph, the network becomes structurally similar to the other topological systems which do not possess geographical constraints. So, the dual representation brings the urban systems outside the geographical field, but in a field of study more complete.

The primale approach is the standard in the creation and the broadcasting of the geospatial data [27]. However, the networks centrality in this geographical representation was not studied as much as those of the social, biological or technical topological networks.

2.2.4 Multiple Centrality Assessment (MCA)

"No matter how good its ofering, mechandising, or customer service, every retail company still has to contend with three critical elements of success: location, location, and location." [36]

The MCA assumes that a good location in the urban environment results from its centrality. A central area is more accessible and this accessibility can be transformed into a better visibility and popularity. A central area will then be characterized by a bigger variety of services, a higher land value or a more extensive use of the ground [28].
The MCA model, set up by the *HSL*, concentrates on the measure of centrality of the urban networks in a **primal** representation.
So the main characteristics of the MCA are:

1. to use a primal approach of the network,

2. to anchor all the measures in the real network of streets,

3. to define the centrality by a set of indexes.

The first two characteristics are the ones which differentiate this approach from the *space syntax*. The last one, differentiates it from the common approaches of network studies in the fields of the regional analyses, transport planning or urban geography. Indeed, the MCA measures of how much a location is central non-only in a closeness meaning, but also how it is intermediate or straight in front of the other locations. All these measures will be detailed further in the theoretical bases (to see: section 3.2, page 20).
PORTA shows that with this tool, better than with the dual approach, we are able to discern the skeleton of an urban structure. The study of the distribution of centralities allows to differentiate auto-organized cities and those with strategic architecture [27]. A case study on the city of Bologna (IT) bring to light strong positive correlations in particular between the centrality of streets and the density of the activities (FIG 2.4) [28].

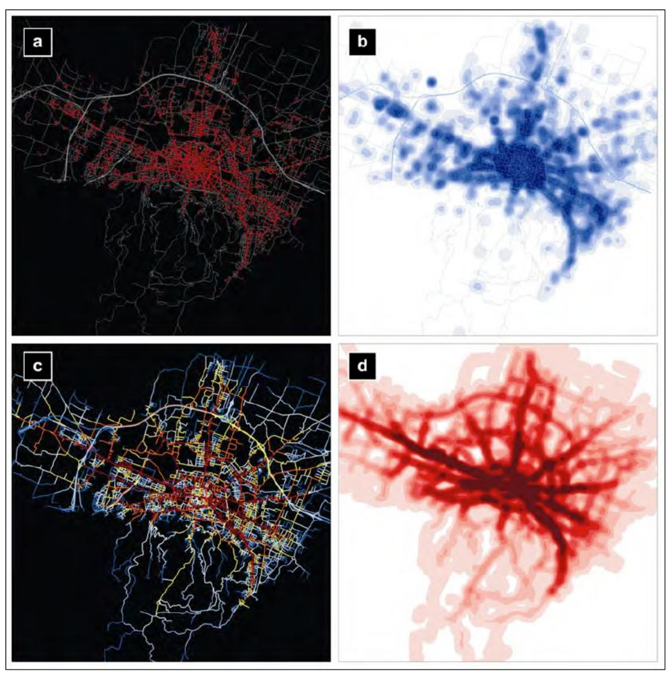


Figure 2.4: Density of activities and street centrality: **(a)** location of retail commerce activities (red dots), **(b)** Density (KDE, h=300m) of retail commerce activities,**(c)** Global Betweenness (C^B_{glob}) of street centrality (blue for lower values and red for higher, **(d)** Density (KDE, h=300m) of C^B_{glo} street centrality [28].

2.3 Kernel density estimation

Within this framework, the Kernel Density Estimation *(KDE) is used to transform the MCA measures and the activities in the same unit which allows a correlation analysis. Furthermore, it is recognized that the function of density is a means to present analyses and illustrations of complex and technical data in a clear and understandable way to the non-mathematicians [33]. The* Network Density Estimation *(NDE) implemented to take into account network distance follows the model of this density estimation.*

In GIS, the KDE is a tool for spatial smoothing and/or for spatial interpolation (Fig 2.5, page 13). Among the methods of *spatial smoothing*[1] and the methods of spatial interpolation [2] the KDE was chosen by Porta in particular because it is one of the *ArcGIS* tools [29, 28].
Anselin uses the KDE in a spatial analysis of the crimes to simplify visually and to examine the complex characteristics of the criminal incidents [1]. Gatrell always in a will of public security, uses the kernel estimator to understand the variation in the risk of disease [19] . Borruso shows that the KDE applied to an urban system, in that case a density analysis of the addresses, allows a better display of the phenomenon but also that it is less sensitive to the size, the position and the orientation than a *Grid Density Estimation* [7].
In an closer field to our, Thurstain-Goodwin applies a KDE to the data of zip codes to obtain a continuous surface of density. Associated with indicators of centrality, the kernel density allows to create a composite urban indicator of centrality applicable to the British cities [37].

The kernel density estimation is a function which balances the events according to their distances. So, two parameters are required. The first one is the *bandwith* (h) which is the distance of influence. Beyond this distance, an event possesses a zero weight. The second parameter is a weighting function (K), mostly it is a normal function. So, Brunsdon proposes an adaptive KDE with a parameter which changes according to the cloud of points structure [10]. He applies also the KDE for a cloud of weighted points [9]. However, if the choice of the parameter h is not coarse, the authors agree that the choice of the function K is less critical [19].

[1] Floating catchment area, kernel density estimates, and empirical bayes estimation.
[2] Trend surface analysis, inverts distance weighted, thin-flat splines, and kriging.

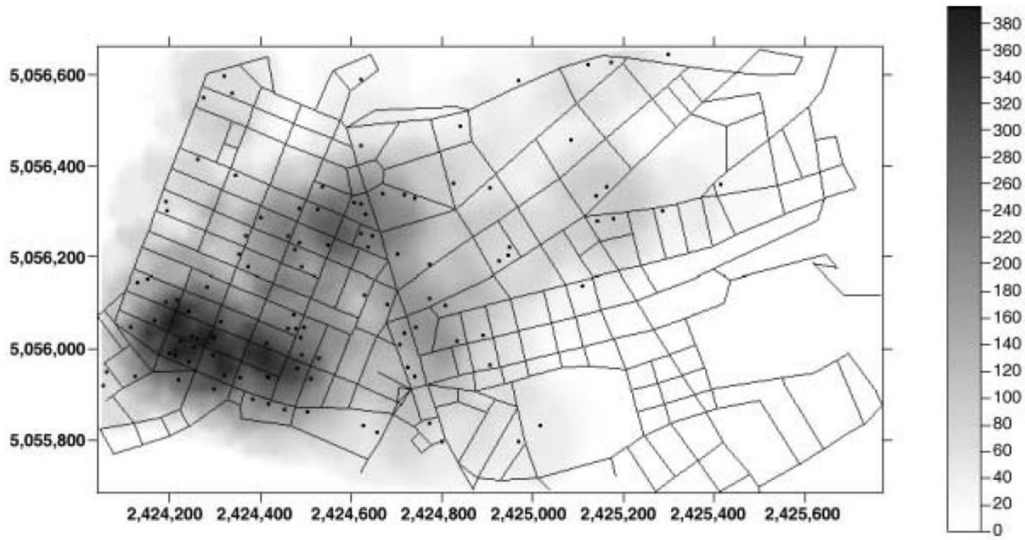


Figure 2.5: Kernel density estimation on bank and insurance branches with a 125m bandwith, Trieste(IT) city [8]

2.4 Network density estimation

The density estimation on the network is one of the tools that proposes this work. This density will be applicable to segments and points balanced by their distance from the point at which is estimated the density.

Miller points out that the hypothesis of a continuous space is too strong in the analysis of the events which take place in a one-dimension sub-space created by a network [23]. In the same way, Batty realizes that the GIS mostly avoid the distortions applied to the euclidean space by the networks constraints. He still points out that if the representation of networks raises no more problem, the next developments which will take place in GIS will have to take into account this constraint. These new point of view will change the perception of the cities development and phenomena which take place there [2].
On this way, an American study illustrates, on the basis of road accident, the risks of statistical use of tools, based on a uniform space, for phenomena forced on a network. A modification of the *K-functions*[3] is proposed to adapt them to networks [40].

[3]These functions allow to analyze the points distribution in particular by taking into account their distance and their intensity.

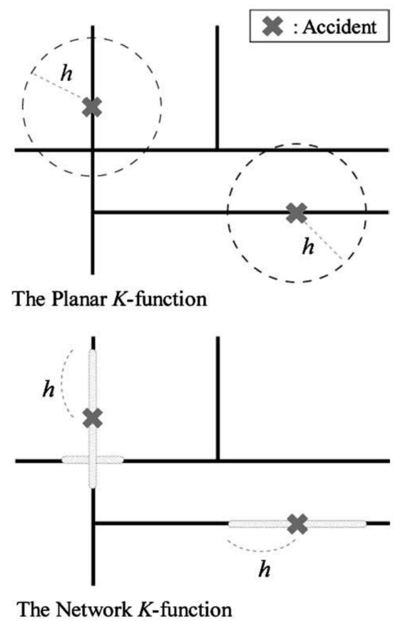


Figure 2.6: Basic concepts of the planar and network K-functions [25]

.

BORRUSO [8] advances that the human activities are forced in the urban network as for example, the addresses refers to streets. However, although these points are points of the network, the methods of density calculation in the GIS use the uniform space and so euclidean distances. He suggests then, to make an analysis of density in the space created from the urban network (FIG 2.7, page 15). He tests then his algorithm on the cities of *Trieste* (IT) and *Swindon* (UK) for the activities related to the branches of banks and insurances. He shows that the difference between the KDE and the NDE (*Network Density Estimation*) is not very important, but also that the NDE allows to better identify linear patterns along the network. However, the density calculated does not take into account a distance weighting function such as one can find in a standard kernel density estimation.

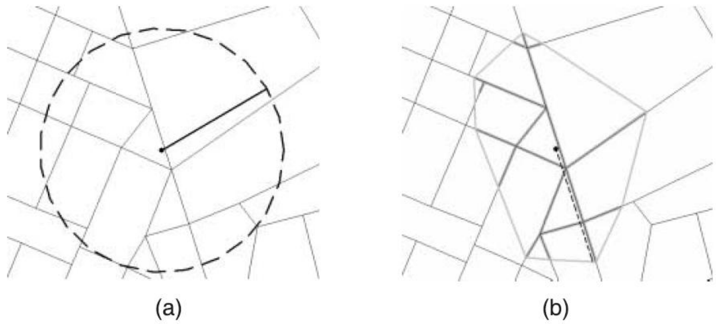


Figure 2.7: Difference between a surface delimited with a euclidean radius **(a)** and a radius applied to the network **(b)** [8]

Finally, OKABE has developed a *kernel density estimation* for estimating the density of points on network. Three kernel functions are proposed, whose two of them are unbiased and which mathematical properties are explained. He applies, this tool for finding "hot spot" of traffic accident. These functions are implemented in a GIS tool and will be at disposal in the extension SANET[4] [24].
Nevertheless, its calculated densities values are attribute of the network and are not generalized to the entire field of the network.

[4] SANET: Spatial Analysis on a NETwork

2.5 Diversity

The diversity in this work, will be applied to six families of activities and service of Barcelona. It will be calculated at once in the euclidean space and in the network space.

The diversity as a measure of richness or community heterogeneousness appears mainly in the fields of the biology and the ecology, namely biodiversity. In other fields, concerning the economy the literature is less important. Nevertheless, the indexes used to measure the diversity in bio-ecology and in socioeconomics share the same principles [26].
In this last category, are classified the studies concerning the town and country planning. During the 20^{th} century, the urban activities were more and more separated in the space. However, they remained connected thanks to the transport revolution. In the urban zones, this separation of the functions becomes a course of action for the territory policy [3]. Today, the tendency goes in the opposite way, creating a diversity in the land-use, which marks a spatial quality for a sustainable economic development [39].

The diversity measure in a raster cell can be derived from equivalent indicators in ecology. A distinction can be made between the **measures of distributions** which indicate the number of species and the distribution of the specimens in these species, and the **measures of variation** which measure the size and the importance of the differences between the present species [39].
For the variety of the services on a network, the first type of indicator which measures the distribution of the specimens is particularly important. One can denote four basic indicators:

- **Richness** m: number of different activities in the area,
- **Dominancy** p_{max}: ratio of the bigest specie in the area,
- **Entropy** H: From the information Shannon and Weaver theory [32],
- **Simpson's diversity index**: represents the probability that two randomly selected individuals in the area belong to the same species.

BATTY uses the measure of richness to describe the multifunctional city (FIG 2.8, page 17). He shows that the multifonctionality depends on the chosen scale. For example, at very fine scale the variety is null, on the other hand, by increasing the size of window it will tend to remain homogeneous in the suburban zones (specialized) but will increase in the city center. In the same study, BATTY introduce the notions of temporal diversity (difference between day and night) and the 3D diverstiy applied to buildings [3].

Figure 2.8: Richness measure in London, low diversity is in blue and high in red: left is the global scale, right is downtown area. [3]

Number of study on the land-use diversity and their relationship with the urban routes appeals to the measure of entropy. BISHOP analyzes the industrial diversity in Great Britain with a measure of entropy in a purpose of town and country planning policy. He points out that the understanding of the specialization dynamics and diversities are critical for the development of the economic urban development policies [5].
Always on the basis of the entropy index, CERVERO makes the hypothesis that the demand in transport possesses three main dimensions: the density, the diversity and the design. On the basis of these '3D', its results support the will of the town planner to create compact, diverse and pedestrian-oriented neighborhoods which can influence significantly the American way of travelling [13].
Very close to this work and with an entropy measure of the land-use, FRANK shows that an increase of diversity in the arrival and departure points of the urban routes tends to decrease the use of *"Single-occupant Vehicle"* [17].

Among all these studies, few of them link centrality (especially in a primal approach) with density and diversity of activity. This is what could be done with the tools that offers this report. To implement this tool it is necessary to go further into the theoretical and mathematical definition of the graph theory, density estimation and diversity.

Chapter 3

Theoretical Basis

In this part, you will find the prerequisite for a good comprehension of the algorithm. First, the graph theory will explain how a network can be represented. Then, you will have a look at some centrality indexes, they are described here to better understand what are the attributes generalized by the estimation of density of the network and what they represent. After that, the kernel density estimation and the diversity will be presented in their mathematical approach.

3.1 Graph theory

A network can be defined under the form of a graph with N *nodes* and K *edges*, as $G = (N, K)$. The nodes n_i belong the set $N = \{n_1 \dots n_N\}$ and the set of the edges linking n_i and n_j is $K = \{n_i n_j\}$, these nodes are *connected* or *adjacent*. In the case of a weighted graph in which every edge has a value (ex: the length), the graph is $G = (N, K, \Omega)$.
The connection matrix of a connected graph without weigth is:

$$R(G) = |r_{ij}|_{n \times n} \text{ with } r_{ij} = \begin{cases} 1 \text{ if } n_i n_j \in K \\ 0 \text{ else} \end{cases} \tag{3.1}$$

The geodesic d_{ij} is the sum of the sides linking i and j in the shortest path between them.

First developped for the study of social network, the centrality measures describe the status of a node in a topological way. The centrality indexes make use of the geodesics to characterize how central is a node.

3.2 Centrality Indexes

The centrality indexes are used in the Multiple Centrality Assessment (MCA) to charaterize the shape of a network using several indexes as the betweeness, closeness or straightness. These indexes are one of the attributes of the edges of the urban network.

3.2.1 Closeness centrality

The proximity of a nodes to the others is the first sensible characteristic in the interaction between nodes. If L_i is defined as the mean of the shortest paths from the node i to the others:

$$L_i = \frac{\sum_{j \in N; j \neq i} d_{ij}}{N-1} \tag{3.2}$$

The closeness index is:

$$C_i^C = L_i^{-1} \tag{3.3}$$

It measures the accessibility of a point, in the way that closer a point to the others, more accessible it is.

3.2.2 Betweenness centrality

Beyond the closeness, the interaction between two distant nodes depends of the nodes belonging the path linking them. These nodes have a strategical location for the control and the influence of the flows. To create this index, only the geodesics linking the two nodes j et k are taken into account. If n_{jk} is the number of geodesics and $n_{jk}(i)$ is the number of geodesic whose i belong, the *betweenness centrality* of the node i is [18]:

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j,k \in N; j \neq k; j,k \neq i} \frac{n_{jk}(i)}{n_{jk}} \tag{3.4}$$

C_i^B takes values between 0 and 1 and reaches the maximum when the node i belong all the geodesics. It measures the volume of flow on a location.

3.2.3 Straightness centrality

This index start from the hypothesis that the communication between two points is better when the path is straight.

$$C_i^S = \frac{(\sum_{j \in N; j \neq i} \frac{d_{ij}^{eucl}}{d_{ij}})}{(N-1)} \tag{3.5}$$

It measure of much the road from i to the other nodes divert from the straight line linking them.

3.2.4 Global and Local index

All these indexes can be computed in a global or local scale, this means that the focus can be done at the entire network or only at a part of it. Local index avoid the *edge effect* which lead to low values in the border of the network [27]. To strike against this effect a local scale (or radius) can be used.

The centrality indexes wich are edges attributes, are generalized in the entire field with a kernel density estimation.

3.3 Kernel Density Estimation

The kernel density estimation (KDE), is implemented in several GIS one of which is ArcGIS. *This density estimation take into account the distance between the events. So, the density is computed for every point (or raster cell) of the map. This continuous field of values allows a visual or statistical analysis between indicators. It has been applied to the activities and centrality weighted edges of Barcelona by the HSL.*

3.3.1 Density Estimation

The probability density function is a basical concept in statistic. For a random event X, its density function f is searched. This function describes the X distribution. The associate probabilities of X are found with the relation [33]:

$$P(a < X < b) = \int_a^b f(x)dx \text{ for every } a < b \tag{3.6}$$

For a set of point of an unknown probability density function, the density estimation is the creation of a density function estimate on the basis of the set of point. There are two approachs. The first, or **parametical** one, looks at the parametrical familly of the function, for example a normal distribution with a mean μ and a variance σ^2. The density function f is obtained on the basis of the estimator for μ and σ^2 and the formulas of a normal density.
For the second or **non parametrical** approach, as the KDE, the data will give the function f without the constraint of a parametrical familly.

The density functions are a tool for the presentation of analysis and illustrations for complex and technical data. The density functions are clear and comprehensible for non mathematicians.

3.3.2 Kernel Estimator

For a set of n events $X_1 \ldots X_n$ for which the estimation of the density is searched, and for $\hat{f}$ the density estimator. The kernel estimator is:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.7}$$

With h being the *bandwith* or the *smoothing parameter*. K is the kernel function which satisfy:

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{3.8}$$

One can see the kernel estimator as the sum of "bumps" at the events. The K function gives the shape of the bump and h set its width (FIG 3.1, page 23). $\hat{f}$ takes the properties of continuity and differentiability of the function K.

3.3.3 2D Kernel Estimator

Multivariate kernel density estimator

The estimator is generalized to the multivariate case, this means for points in a d-**dimensions space**. In this case, the kernel density estimator is defined:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.9}$$

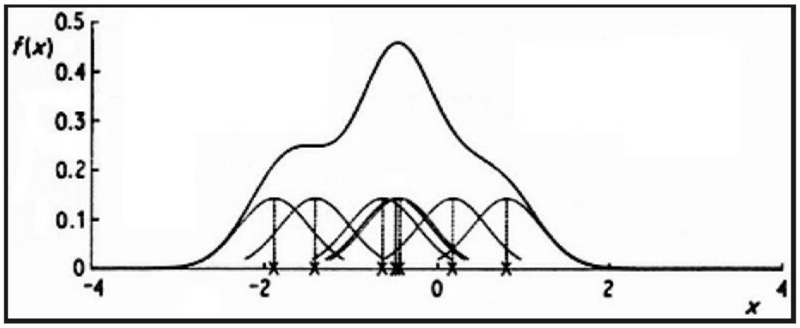


Figure 3.1: Estimation de Kernel avec les Kernels individuels pour $h = 0.4$ [33]

This time, the kernel function is defined for x in a d-dimensions space:

$$\int_{R^d} K(x)dx = 1 \tag{3.10}$$

Often, the K function is a radially symmetric unimodal probability density function as the standard multivariate normal density function[1]:

$$K(x) = (2\pi)^{-d/2} exp(-\frac{1}{2}x^2) \tag{3.11}$$

An other multivariate kernel function is the one from Epanechnikov[2]:

$$K_e(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-x^2)^d \text{ if } x^2 < 1 \\ 0 \text{ else} \end{cases} \tag{3.12}$$

Where c_d is the volume of the unit d-dimensional sphere: $c_1 = 2$, $c_2 = \pi$, $c_3 = 4\pi/3$, etc.
Thus, if $d = 1$:

$$K_e(x) = \begin{cases} \frac{3}{4}(1-x^2) \text{ if } x^2 < 1 \\ 0 \text{ else} \end{cases} \tag{3.13}$$

Finally, an other useful[3] kernel for $d = 2$ [4]:

$$K(x) = \begin{cases} \frac{1}{3\pi}(1-x^2)^2 \text{ if } x^2 < 1 \\ 0 \text{ else} \end{cases} \tag{3.14}$$

[1]It is used in Crimestat a spatial statistics program for the analysis of crime incident location.
[2]Silverman, 1986, p. 76, equation 4.5 [33].
[3]The Epanechnikov kernel function is modified for computational considerations.
[4]The reference of the ArcGIS KDE goes to this function.

Spatial kernel density estimator

To create an estimate of the density for a cloud of points, a sliding window over the data is used. An evaluation of the density is done in this window. For the kernel density estimation, the window is a $3D$ sliding function which weights the events according to their distance from the point on which the density is evaluated. In a more formal definition with s a location vector over the field R and $s_1 \dots s_n$ the location vectors for the n events. The intensity estimation $\lambda(s)$ in s is :

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} K\left(\frac{s - s_i}{\tau}\right) \tag{3.15}$$

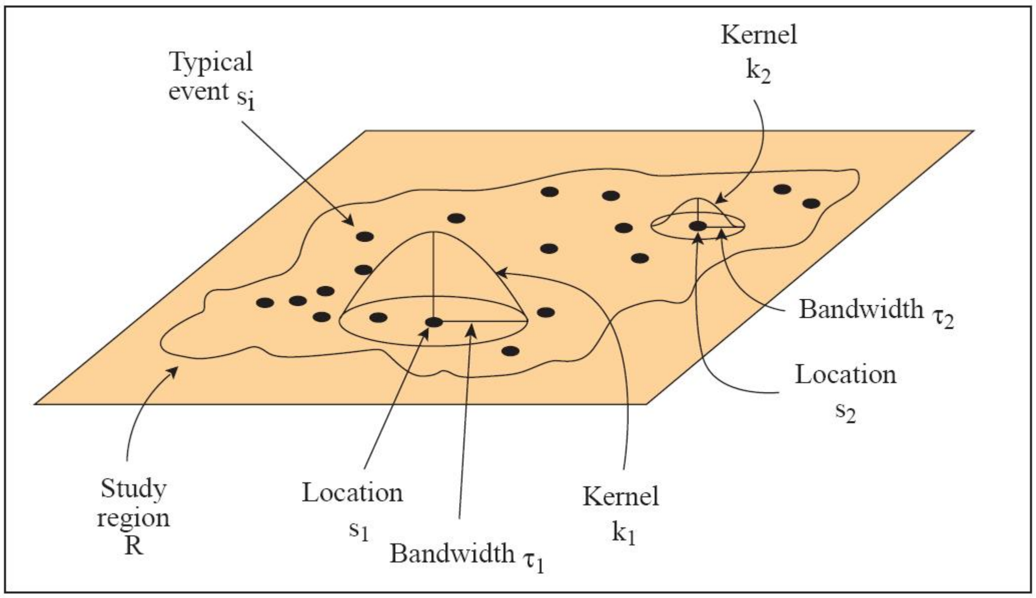


Figure 3.2: Kernel density estimation of a cloud of points [1]

The K function is a $3D$ function which goes in each point r of R, (Fig 3.2). The distances of the events s_i which belong the bandwith (here τ) are computed and contribute to the intensity estimation in R according to their proximity.

Several authors find that among the K function, the differences on the results are not significant[16]. However, the h choice is not so coarse. Some suggest to optimize h according to the structure of the cloud [11] or to change it according to the point distribution[9].

Several kernel function are implemanted in the GIS[5]. *ArcGIS Spatial Analyst* has a function for the points and the lines, but allows the use of only one kernel function, known as *quadratic* or *Epanechnikov* which was presented earlier in the equation 3.14:

$$K(x) = \begin{cases} \frac{1}{3\pi}(1-t^2)^2 \text{ if } t^2 < 1 \\ 0 \text{ else} \end{cases} \tag{3.16}$$

With $t = d_{ij}/h$. The value in each grid point j in a distance d_{ij} of the event i is obtained from the sum of the individual kernel functions of the point belonging the bandwith.

The distance used in the kernel density function is measured in the euclidean space. For the network density estimation this euclidean distance will be replaced by a network distance find thanks to Dijkstra.

3.4 Shortest Path Algorithm: Dijkstra

This shortest path algorithm, in its common use, allows to find the succession of nodes on the shortest path between two points. In this work, it is used to create a shortest path tree of a defined length. The densities and diversities will be computed on this tree.

One of the widespread algorithm to find the shortest path from a point to an other is *Dijkstra* [15]. It is used for a graph whose edges values are positive or null. It is easy to implement and its comprehension is simple, so it is good algorithm for the urban networks in their graphical representation (Fig 3.3, page 26).
Its output is two values for each node. The first one is the length (or total weight) of the shortest path. The second one is the previous node. With this two values, it is possible to select the node belonging a fixed radius and to create the shortest path finding the successive previous nodes.

Once the shortest path tree is created, the parameters as the densities or the diversities can be computed.

[5] www.spatialanalysisonline.com

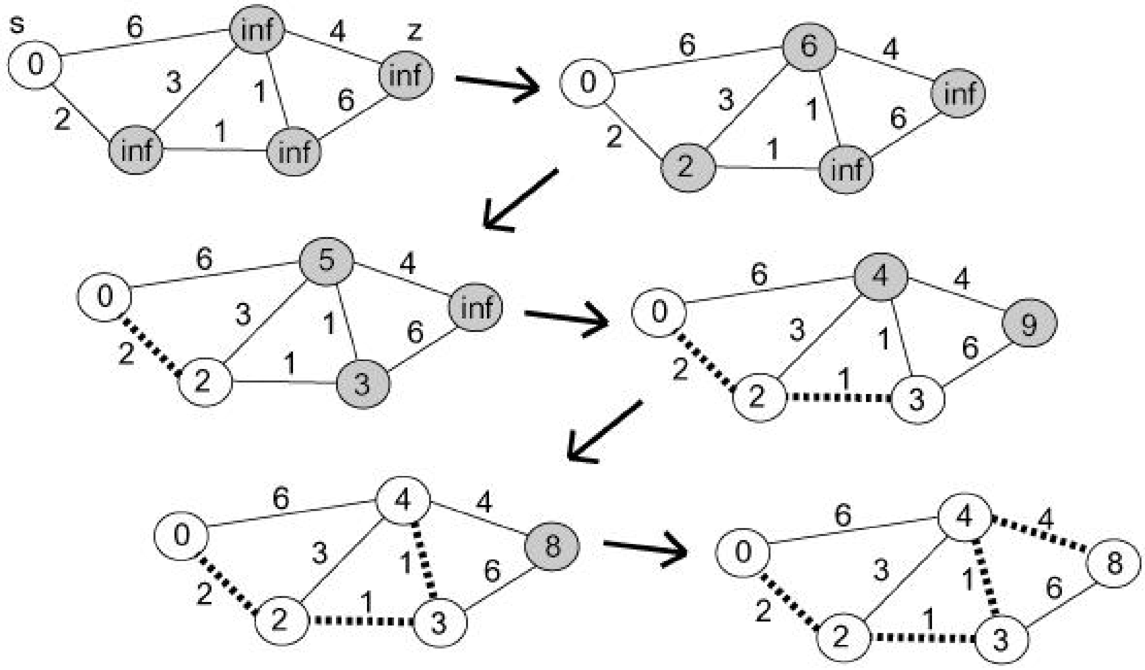


Figure 3.3: Dijkstra's algorithm: the nodes are initialized with a *infinite* value. Then each node is vistited begining with the nearest one. The shortest path and shortest distance are kept (www.heptargon.de)

3.5 Diversity

A mesure of entropy or Shannon's diversity will be applied to the activities of Barcelona. According to many land-use and land-planning studies, our algorithm will compute an entropy index in both the euclidean and network space. Here are some diversity indexes which will help to understand what diversity is.

A measure of diversity inside a cell can be derived from equivalent indexes in ecology. A distinction can be done between the measures of distribution which look at the number of species and the distribution of the specimens inside the species, and the variation measures which look at the size and ratio between the species[39]. With a focus on the diversity of activities on a network, the first familly of indice measuring the distribution are particulary pertinent.
There are four basis index:

- **Richness** m: number of different species in the field;

- **Dominancy** p_{max}: ratio of the biggest specie;
- **Entropy** H: From the information theory of SHANNON and WEAVER [32];
- **Simpson's diversity index**: represents the probability that two randomly selected specimens in the area belong to the same species.

For the next sections, it is even necessary to define: S the number of species, i the different species, n_i the number of specimens in the specie i and N the total number of specimens.

3.5.1 Measure of richness

The measure of richness is simply the number of different species inside the field. This measure is easy to compute but is light in information. It is often use before the complete study. It does not take into account the uniformity in the distribution inside the field. It means that it dont give informations about the ratio of a specie.

$$m = S \tag{3.17}$$

One can find some variation of this richness index.

3.5.2 Dominancy

It is known as diversity index from Berger-Parker as well. With $p_i = n_i/N$

$$p_{max} = max_{1<i<N} p_i \tag{3.18}$$

3.5.3 Simpson's diversity index

In sociology and psychology, this index is known as Blau's index. It takes into account, in the same time, the number of species and their relative abundance. It represents the probability that two specimens randomly choosen belong to the same specie [34].

$$D = 1 - \sum_{i=1}^{S} p_i^2 \tag{3.19}$$

For a little number of specimens, it can give to high results in the areas with low diversity.

3.5.4 Entropy index

The Shannon's diversity index or communication entropy is the index of diversity the most used. Its mathematical basis moreover are the most robust. It is used in ecology, computer sciences and telecommunication.

$$H = -\sum_{i=1}^{N} p_i \ln p_i \tag{3.20}$$

The entropy index takes the lower values, when there is only one specie in the sample. In contrary, it takes the highest values when there are a lot of species with a same ratio.

With these detailed knowledges of the the tools which will be the basis of the indexes developed in this report, the algorithm will be presented in the next part.

Chapter 4

Development of the tool

This chapter first presents the algorithm leading to the calculation of the network densities and diversities. Then, a focus on the support of these calculations will be made. So, the constraints and choices and then the algorithm with a more computer sciences approach are explained.

4.1 Algorithm

In this part, you will find the big steps of the algorithm allowing the calculation of the densities and diversities. If you need more informations for a step you will find it on the focused algorithm which explains in ArcGIS and PostGIS vocabulary what is really done by the script.

4.1.1 Inputs and Goals

The feed of our algorithm is two files. The first one is the network with its values of centralities. With the tool that HSL has developed, one can generate a topologicaly clean network as well as its nodes. The edge's attributes of the network are the values of the *from/to nodes*, their coordinates, the length, and several values of centrality. The other file contains the activities with their class (Fig 4.1, page 30).

The goal of the tool is to compute the *density of centralities* on the network, the *density of activities* on the network, the *diversity of activities* in an euclidean space as well as the *diversity of activities* on the network. All the indexes dealing with the network will be based on a shortest path tree (SPT). The shortest path tree is the collection of all the segments reachable within a given distance (Fig 4.2, page 30).

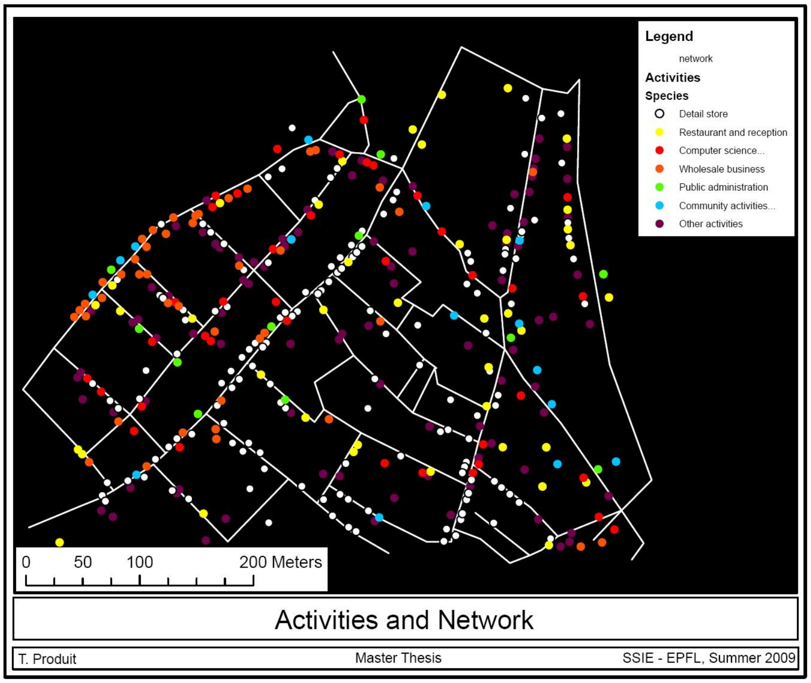


Figure 4.1: Activities and network for an extract of Barcelona

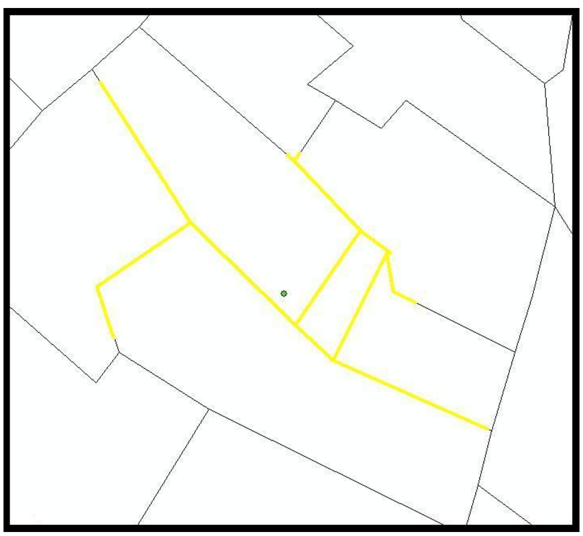


Figure 4.2: The 100m shortest path tree for the green point

4.1.2 Generation of a raster

The first step is the generation of a raster over the network. This raster is void at first, but it will receive the different values at the end of the algorithm. The centroids are calculated, they will be the start point of the SPTs (Fig 4.3).

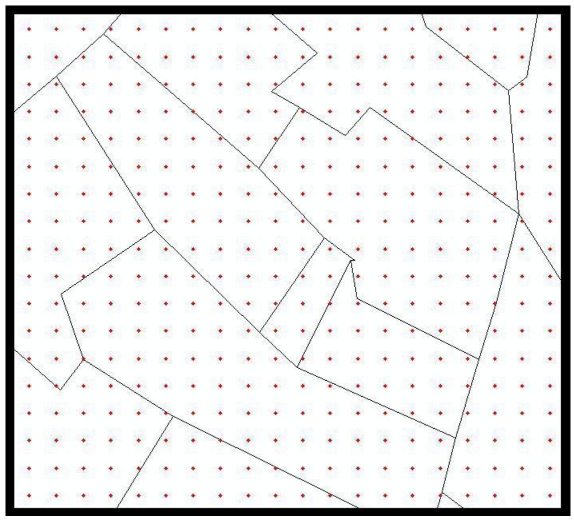


Figure 4.3: The 10m grid of point

4.1.3 Clip of the neighbor network

In the next steps, the shortest path algorithm will allow to create the SPTs. It would be too time consuming to apply this algorithm on the whole network. The trick used here is to clip the network inside a defined radius around the centroids. The clip has as radius the bandwith value, the clip function keeps the edges touching the circle (Fig 4.4, page 32).

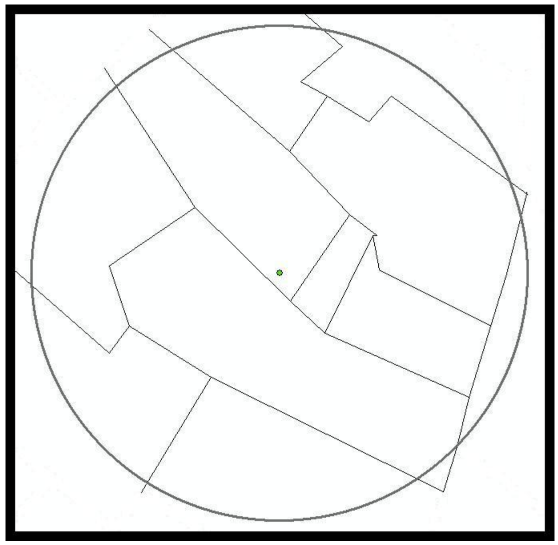


Figure 4.4: The 100m clipped network for the green grid point

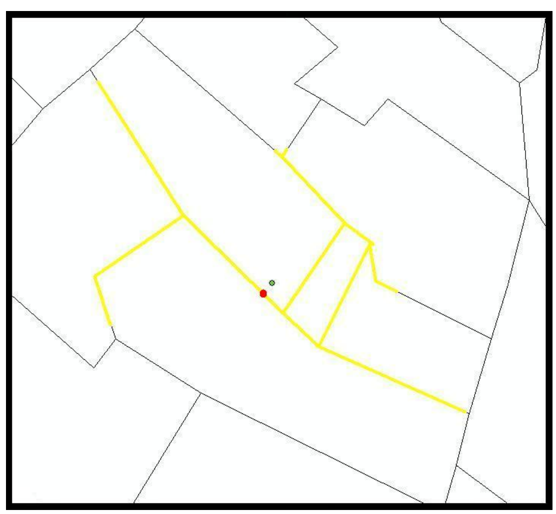


Figure 4.5: The 100m shortest path tree for the green grid point, the start point is red

4.1.4 Projection of the centroids

The centroid is projected on the nearest segment of the network (Fig 4.5, page 32). This segment is cut into two part at the projection location to allow the start of the shortest path algorithm in both directions. The distance between the centroid and its projection is kept in memory.

4.1.5 Shortest path algorithm

Here Dijkstra has been used. The real start point of the algorithm is the projection of the grid point on the edge. However, a virtual edge is created between the grid point and the start point to take into account this distance. The results of this algorithm are the distance from the centroid to every node as well as the successive nodes creating the shortest path.
This algorithm allows the selection of the segment belonging the bandwith (Fig 4.5, page 32). Nevertheless, some changes to the standard Dijkstra have been done to perform the needs of this work.

4.1.6 Network density

The network density is created with a focus in the *kernel density estimation*. This means that the events receive a weight according to their distances as in a KDE.
At this point, two choices are made. First, the dimension d of the space, indeed the events are distributed in a 2D space, but the measures are done in a 1D space delimitated by the network. Differently, the results can be a sort of linear density or a sort of surface density. Our goal is to be close as possible from the *ArcGIS* kernel density function, thus the surface approach is applied.
The second choice is the shape of the K function, with this same previous goal, our choice is the modified Epanechnikov function which is quoted by *ArcGIS* (to see: section 3.3.3, equation 3.12, page 23).
In this way, for the *network density*, an Epanechnikov 2D kernel density is applied to a non-uniform space created by the shortest path tree.

Density of activities

The activities are projected on the shortest path tree and the network distance from the grid point to every activity is calculated. As in a *kernel density*, a weighting function is applied to these distances (Fig 4.6). The sum of these values gives the density of activity (to see: section 3.3.2, page 22).

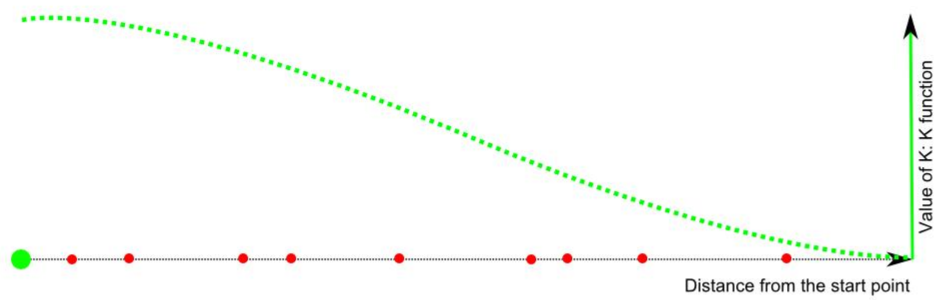


Figure 4.6: NDE of the activities, the weights for the activities (red dots) are represented by the green curve (K function)

Density of network centrality

Here, for every segment of the SPT, the distance from the grid point to the middle of the segment is calculated. This distance is the input for the weighting function. Its output is multiplied by the length of the edge and the value of centrality which are the *population*. The length is used to take into account that long edges are more important than short ones (Fig 4.7).

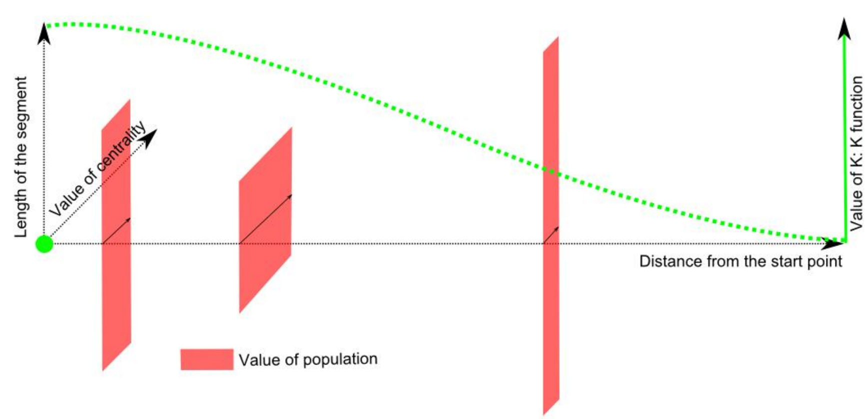


Figure 4.7: NDE of the network, the weight for the edge are represented by the green curve (K function), the edge becomes a value of population, which is the length time the value of centrality.

4.1.7 Euclidean diversity

For every centroid the activities within a given radius are clipped (Fig 4.8). The diversity is calculated with the Shannon's formula (to see: section: 3.5.4, page 28).

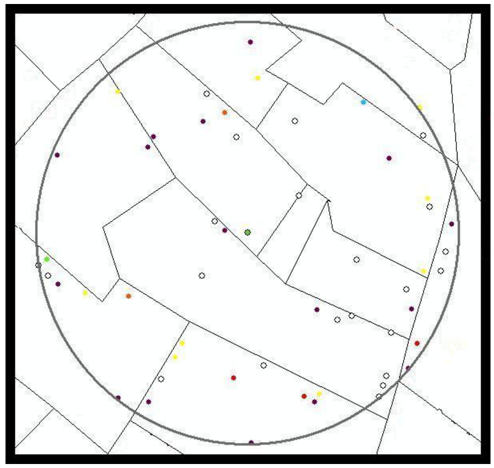


Figure 4.8: Clipped activities for a bandwith of 100m in the euclidean space

4.1.8 Diversity on the network

The diversity is calculated for the activities projected onto the shortest path tree (Fig 4.9). For every centroid, the activities within a given radius are clipped. The diversity is calculated with Shannon's formula.

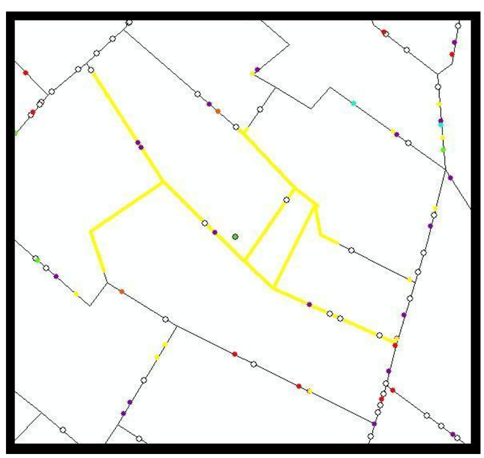


Figure 4.9: Activities and the 100m SPT

4.2 Constraints

To integrate the previous algorithm in a real script, it is necessary to condider the constraints and choices which lead to the tool that provides this work.

4.2.1 ArcGIS

First, the *Human Space Lab* is working with *ArcGIS 9.1* in its normal configuration, namely *ArcView*. Knowing (and testing) that the downgrade is difficult between versions, the development is done in *ArcGIS 9.1*. At first, the algorithm was very *ArcGIS*-dependant. Now, for several reason this program is used essentially for display and rasters management.

4.2.2 Large data

The tool will be used on networks of big cities like Barcelona. For this city, there are more than 11,000 edges and 166,000 activities. Our tool has to deal with such data. The next constraint ensues directly from this one.

4.2.3 Time

As you can imagine, the ressources needed to create a shortest path tree for every point of a raster cell are important and the time needed to complete this process as well. The choice of time scale is not trivial: what is acceptable and what is not? Probably a year is too much (which was the first solution proposed), but one month? One week?
The faster the results come, the more the tool will be used and researches make progress. The principal goal is to reach a satisfactory computing time. And if possible, this time should allow to obtain results before the deadline of this report[1].

[1] The last version compute the parameters in around 30 hours.

4.3 Choices

4.3.1 Language

Several languages can be used to script within *ArcGIS*. Moreover, a lot of examples are given in *Python scripts* in the *ArcGIS Help*. This language has a lot of free extensions. Some of them allow the connection to a *PostgreSQL* database. So, naturally the choice is gone in this way. *Python* is an object-oriented language that is quite easy to use without previous knowledge. Moreover, the *model builder* of *ArcGIS* gives the possibility to translate a model in this language, this ability enables to easily begin understanding the language as implemented in *ArcGIS*.

4.3.2 Spatial calculations

In its first version, the algorithm proposed, used tools from the standard *ArcToolbox*. Though all calculation are possible with these tools, the required time was very long. For an extract of the network of about 8.5 km^2, the calculation would take more than 90 days! After this observation, the choice was made to realize the spatial calculations in a *spatial database*.
In the world of the spatial database, a first distinction can be done between *open source* databases and the others. Obviously, an open source database which will allow more development and is easily accessible for everyone, was choosen.

The database *PostgreSQL* after 15 years of active development is now recognized for its reliability and integrity. *PostgreSQL* is the foundation for an other project: *PostGIS*. This last one adds support for geographic objects, allowing it to be used as a spatial database for geographic information systems (GIS), much like ESRI's SDE or Oracle's Spatial extension. *PostGIS* has two very useful functions:*shp2pgsql* and *pgsql2shp* to convert shape files into databases and inversely.

4.3.3 Access to the database

Several *DB API 2.0 Drivers* are suggested by *PostgreSQL*[2]. The choice was done for the one which permits an access to *PostgreSQL 8.3* and works with *Python 2.1*.

[2] http://wiki.python.org

So, for this project the API *psycopg2*[3] was used. To work well, this API needs the installation of the *eGenix.com mx Base Distribution*[4] which is free as well.

All this material is the computer basis of the algorithm, which is detailed in the following pages.

4.4 Flowchart

In the FIG 4.10, page 39, the flowchart of the data is represented. *ArcGIS* is simply used to prepare the data and for the management of the raster. All the calculations are done between *Python* and *PostGIS*: graph and mathematical calculations use *Python*, spatial request call for *PostGIS*.

[3] http://initd.org/projects/psycopg2
[4] http://egenix.com

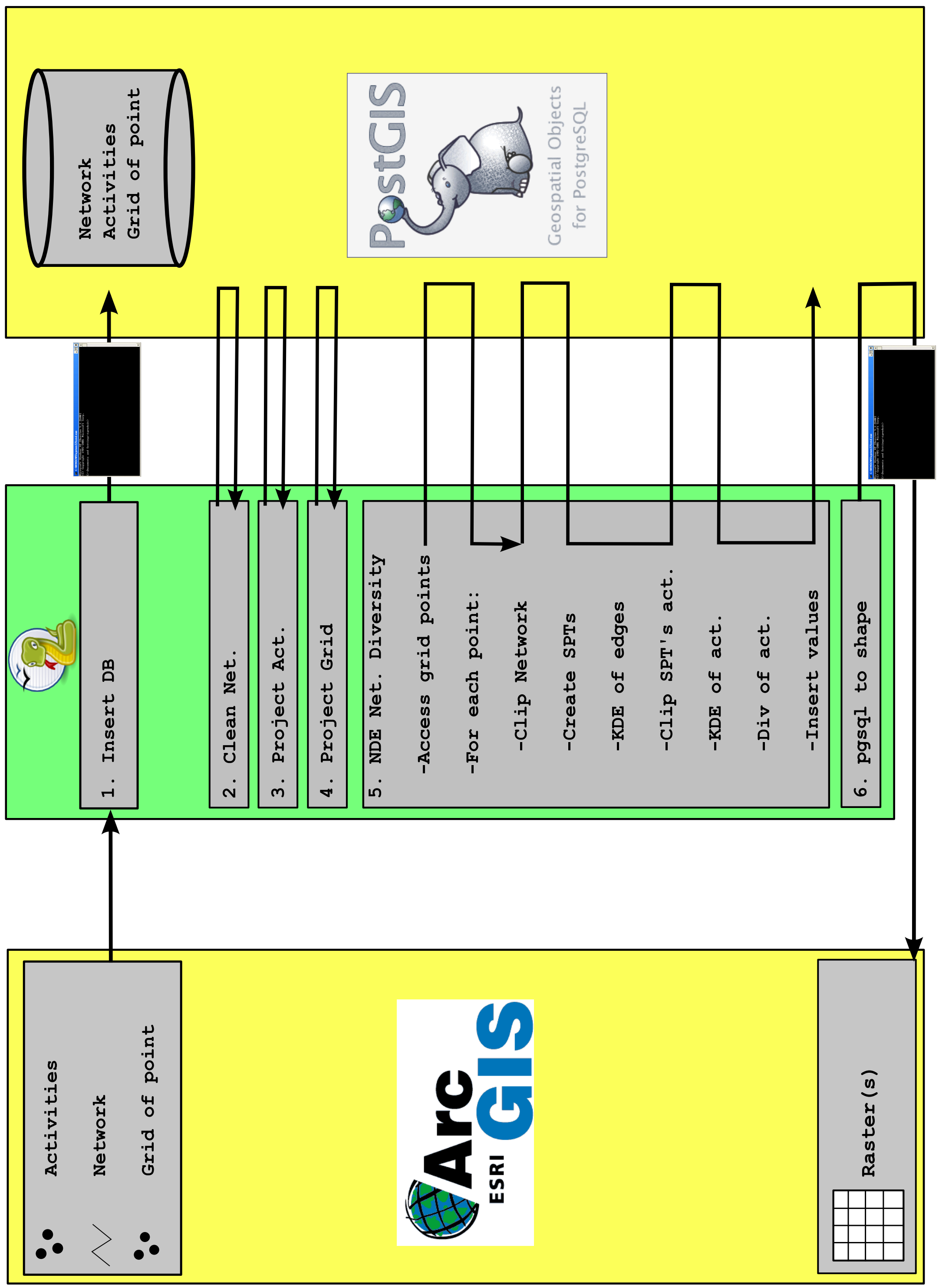


Figure 4.10: From the shape to the rasters of NDE values.

4.5 Detailed Algorithm

In the following sections, the algorithm is explained more accurately with a focus on the functions and tools used.

4.5.1 Generation of a raster

The first step is the generation of a (constant) raster over the network. Thus, the function *Create Constant Raster*[5] in *ArcGIS* is used.
Then, the centroids of every cell of the raster grid are found with the *ArcGIS* function *Raster to Point*[6]. At this time, a grid of points over the network as a shape file is available.

4.5.2 Shape to database

The activities, the network and the grid of points are exported to the database with the use of *shp2pgsql.exe*[7] which is a function of *PostGIS*.

4.5.3 Projection of the centroids onto the network

The goal of this step is to find for every centroid its nearest edge and the distance from the centroid to this edge. At present, the function of the nearest neighbor is not available in *PostGIS*. Therefore, this search is done in force, meaning that for a choosen radius (*ST_Dwithin*[8]) the algorithm looks at the distance to every edge (*ST_Distance*[9]) to keep the nearest. The request is done in *PostGIS*, extending the table of the centroids with their nearest edge and the distance.

4.5.4 Locate grid point on the network

With the previous step, for each grid point, the nearest edge and its distance is known. Now, the coordinates and geometry of the point on the edge and its linear measure

[5]Spatial Analyst Tools/Raster Creation/Create Constant Raster
[6]Conversion Tools/From Raster/Raster to Point
[7]C:/Program Files/PostgreSQL/8.3/bin
[8]ST_Dwithin PostGIS documentation
[9]ST_Distance PostGIS documentation

have to be found (the projection is at 50m from the beginning of the edge). To do so, first the function *ST_Line_Locate_Point*[10] allows to find the ratio of the edge at which occurs the projection. Then, the function *ST_Line_Interpolate_Point*[11] allows to access to the coordinates of the projection (Fig 4.11).

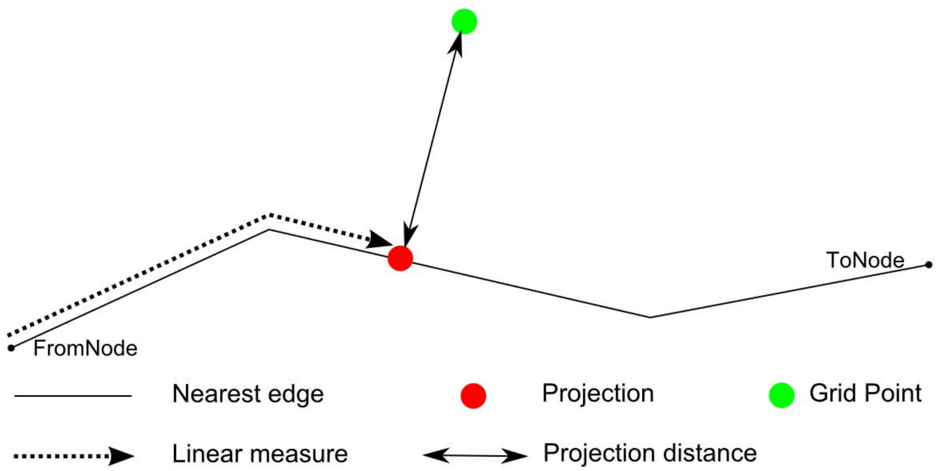


Figure 4.11: Projection of a grid point on the nearest edge of the network

4.5.5 Clip of the neighbouring network

For every point of the grid, the selection of the surrounding network is done with the idea of applying the shortest path algorithm only on this substract. First, a buffer around the centroid is created with the *PostGIS* function *ST_Buffer*[12]. Next, the clip of the network inside this disc is done with the function *ST_Intersects*[13]. This function keeps the edges belonging or touching the disc.

4.5.6 Shortest path algorithm

Before using Dijkstra as shortest path algorithm, the graph of the clipped network has to be built in *Python*. Several informations are needed: the edge's ID, the *from*

[10] ST_Line_Locate_Point PostGIS documentation
[11] ST_Line_Interpolate_Point PostGIS documentation
[12] ST_Buffer PostGIS documentation
[13] ST_Intersects PostGIS documentation

node's ID, the *to* node's ID and the lengths of the edges. All these values are allready written in the table of the clipped network.
The graph needed for Dijkstra should have the form:

- Node$_i$ has as neighbors (N_a, ..., N_z)
- Pair(N_i, N_j) has the weight L_{ij} (here the length)

To create the list of the neighbors for every node of the graph, *Python* has an interessant data structure: the *dictionary*. The dictionary allows to put in memory a *list* of items for a given *key*. For our problematic, the *key* is the current node, and the *list* is its neighbors.

Start Point:

The start point of the shortest path algorithm is the projection of the centroid onto its nearest edge. In this way, two new edges are created: from the start point to the the first extremity and from the start point to the second extremity of the edge (Fig 4.11, page 41). For both of these new edges, the new lengths are calculated and inserted as weights. The value of the distance at the start point is taken as the distance from the centroid to this point.

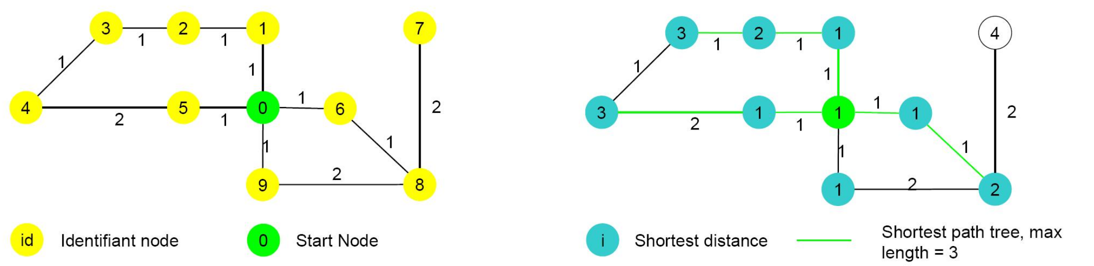
Results:

The results from the Dijkstra algorithm are two *Python dictionaries*. The first one contains all the nodes with the value of shortest distance, the second one contains all nodes and their previous node. With these results, one can find the path to every node for which the weight doesn't overpass the choosen bandwith.

Modification:

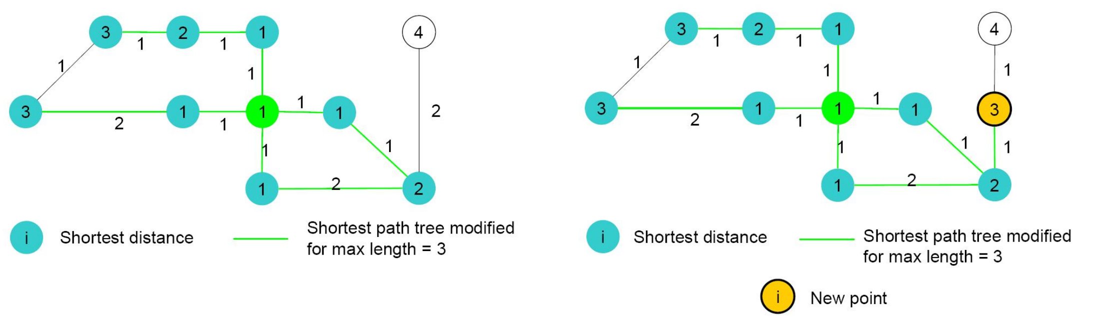
Two modification are done to the "normal" Dijkstra algorithm:

- First, all the edges which are crossed with a value of distance that doesn't go beyond the choosen bandwith are put in the tree.
- Then, the shortest path tree is completed to reach the choosen distance (Fig 4.12, page 43).



(a) Original network, every edge has a weight and in the circle are the ID's of the nodes

(b) Dijkstra's results: every node has as attribute the weight of its shortest path, the tree is in green



(c) First modification: the edges which are crossed with a distance shorter than the bandwith are added to the tree

(d) Second modification: the tree is completed to reach the bandwith

Figure 4.12: Steps leading to the shortest path tree

4.5.7 Network density of activities

To find the coordinates of the activities on the network, their distances and their measures, the same steps than for the grid point are followed (see sections: 4.5.3 and 4.5.4, page 40).
All the activities are first projected onto the network. The result of this projection is the ID of the edge, as well as the distance from the projected activity to the edge's first node (in the drawing meaning). Thus, a value of distance from the activity up to the start point can be calculated. This distance is the input (x) of the *kernel*

density function. First, x is normalized by the bandwith h. So $t = x/h$.

$$K(t) = \begin{cases} \frac{1}{3\pi}\left(1-t^2\right)^2 \text{if } |t| \leq 1 \\ 0 \text{ else} \end{cases} \tag{4.1}$$

And then:

$$\hat{f}(x) = \frac{1}{n \cdot h^2} \sum_{i=1}^{n} K(t) \tag{4.2}$$

4.5.8 Network density of street centrality

The network density of street centrality is the density of the street with a value of population made from the centrality (C_i) and the length (L_i). The population values determines the the number of times to count the edge. For every *shortest path tree* edge, the distance from its center of gravity to the start point is calculated. This distance is the input (x) of the *kernel density function*. x is normalized by the bandwith h. So $t = x/h$.

$$K(t) = \left\{ \begin{array}{l} \frac{1}{3\pi}\left(1-t^2\right)^2 \text{if } |t| \leq 1 \\ 0 \text{ else} \end{array} \right. \quad (4.3)$$

And then:

$$\hat{f}(x) = \frac{1}{h^2 \cdot \sum_{i=1}^{n}[L_i \cdot C_i]} \sum_{i=1}^{n}[K_i(t) \cdot L_i \cdot C_i] \tag{4.4}$$

The length of the edge is used to give more weigth to the long edges than to the short ones.

4.5.9 Euclidean diversity of activities

For the euclidean diversity, first a bandwith is chosen, then in this radius all the activities are clipped. The activities were divided into species. In this way, they have an attribute (integer) representing their species. After that, Shannon's diversity index is applied (to see: section: 3.5.4, page 28), namely for every species in the disc, the specimens are counted and then their ratio (p_i) computed. Next, the following equation can be applied to find H.

$$H = -\sum_{i=1}^{N} p_i \ln p_i \tag{4.5}$$

4.5.10 Network diversity of activities

To compute the network diversity, the only difference is that the projected activities onto the network belonging the SPT are used.

Chapter 5

Case Studies

5.1 Extract

This first case study is an extract of the Barcelona's network. Its goal is to understand the results that one can expect of the algorithm. The advantage of this scale is a good visualization of activities and raster cells. The disadvantage is an important edges effect. The presented parameters are: the *total length* of the shortest path tree, the *number of activities*, the *kernel* and *network density* of **activities**, the *kernel* and *network density* of the **edges**, the *kernel* and *network density* of **edges centralities**. All these parameters are illustrated with maps.

5.1.1 Context

Now that the algorithm is understood, a case study will show the results. An extract (500m x 500m) of the Barcelona's network was chosen. For this network, values of centralities are not recomputed. Only those of the whole network are used. In order to better understand the results, only one activity by location is kept. This means that if some different activities are in the same building, only one appears in the calculation and in the maps. This will allow to better understand the density and diversity of activities.
This extract is to small to be representative of the results of a whole network. Edge effects, especially for the density of network, can lead to "strange" patterns. However, with this scale a good visualization of what the algorithm does, and what a shortest path three looks like is possible.
If you are used to the raster results of *kernel density*, remember that this function gives a good display of the results, with smooth patterns on the map. Don't be

surprised with more granular raster pattern for the *network density* results. They reflect the projection of the centroid on the single nearest edge rather than on all the edges around it.

For a good display and comprehension, the classification colors for the rasters go from white (low values) to black (high value), the other parameters added to the maps are in color. Finally, the classification is always done with the *quantile* (same number of values in each class) of six classes.
First, some preliminary values which will help to understand the next parameters as the network density of activities and centralities are presented. Then, the focus will be made on the euclidean and network diversity.

5.1.2 Total length of the Shortest Path Tree

Remember that for each centroid of the raser cell a shortest path tree is created. The map in Fig 5.1, use a bandwith of 100m for the SPT. Here the total length of the SPTs is showed. This indicator represent the density of the network. The denser the network, the longer the SPT will be.
The following figure shows three examples of SPT: the yellow one measures 427m, the blue one 388m and the orange 238m. These values represent the sum of the lengths of the colored segments. So, near long edges, as on the right side of the map, this value is always near 200m (100m going upwards, 100m going downwards). In locations where the density of edges (or density of crossroads) are high, the total length increases.

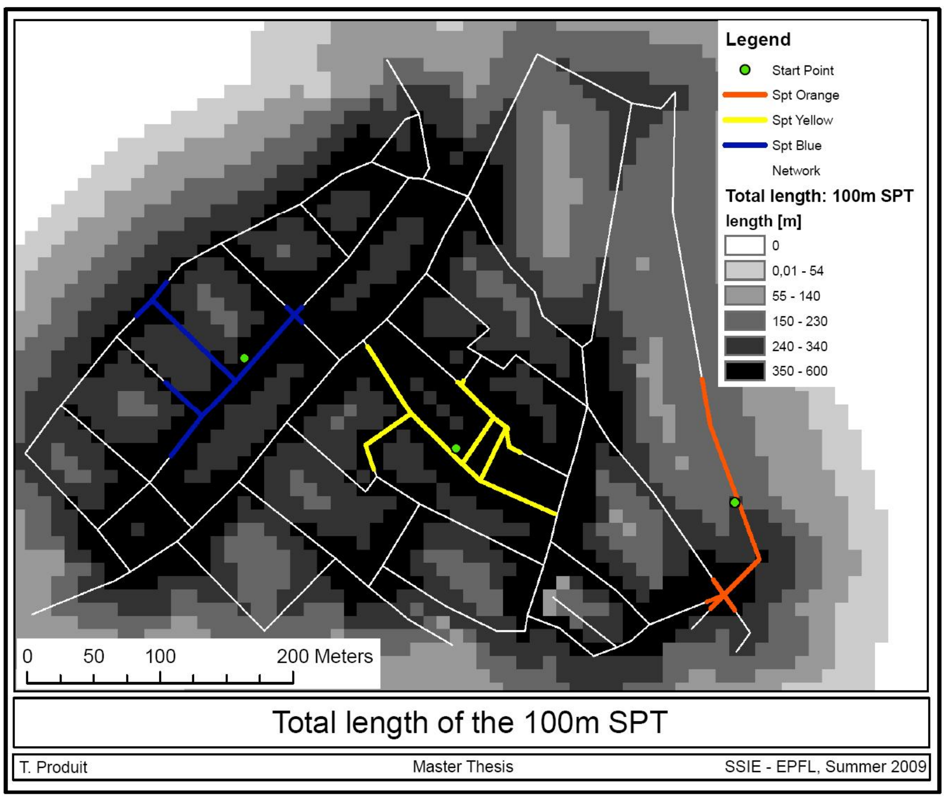


Figure 5.1: Total length of the SPT with a 100m bandwith

5.1.3 Number of activities

In the map of the Fig 5.2, the raster exhibits the number of activities for the SPTs. This indicator can be understood as a mix of the density of the activities and the density of the network. In the areas where the network is dense, the SPTs are long and are able to detect a lot of activities. Nevertheless, if there is no activity on the SPT, the value will stay low.
In the following map, three SPTs with their start point are given for example. There are 45 activities on the blue one 14 in the yellow and 0 in the orange.

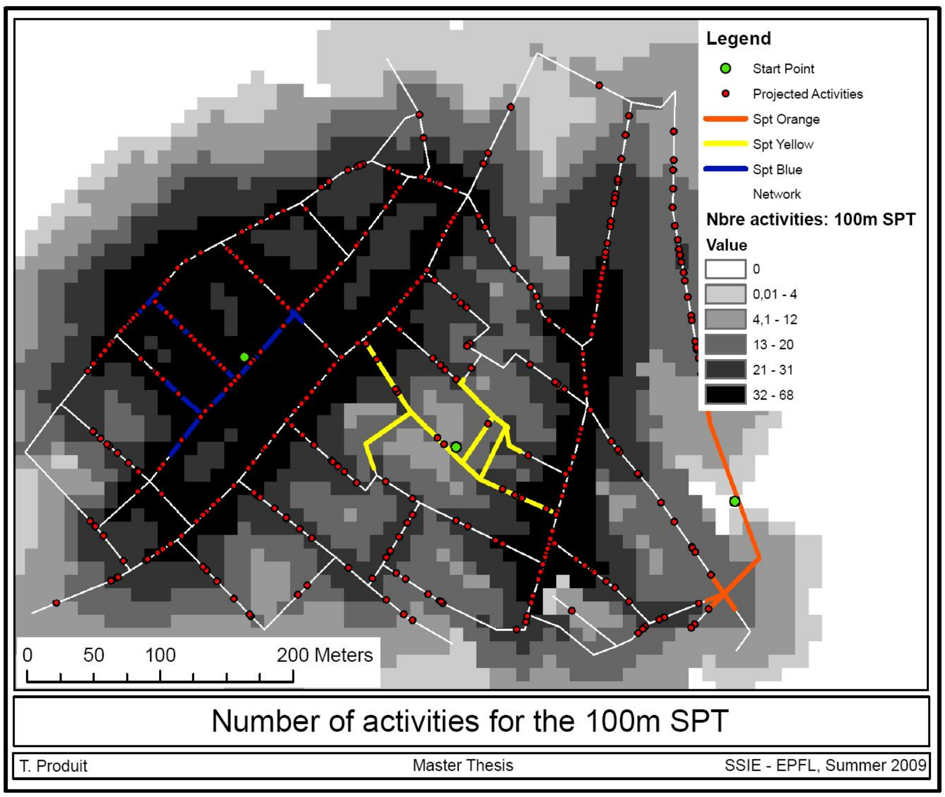


Figure 5.2: Number of activities with a 100m bandwith

5.1.4 Kernel density of activities

For the map of Fig 5.3, the standard *kernel density* of *ArcGIS* is applied to the activities for a 100m radius. Three clouds highlight the locations with a high density of activities. The patterns are smooth all over the map.
One can see that this map and the previous one (Fig 5.2, page 49) show high values at the same location. However, the number of activities on the SPTs focus on the network edges.

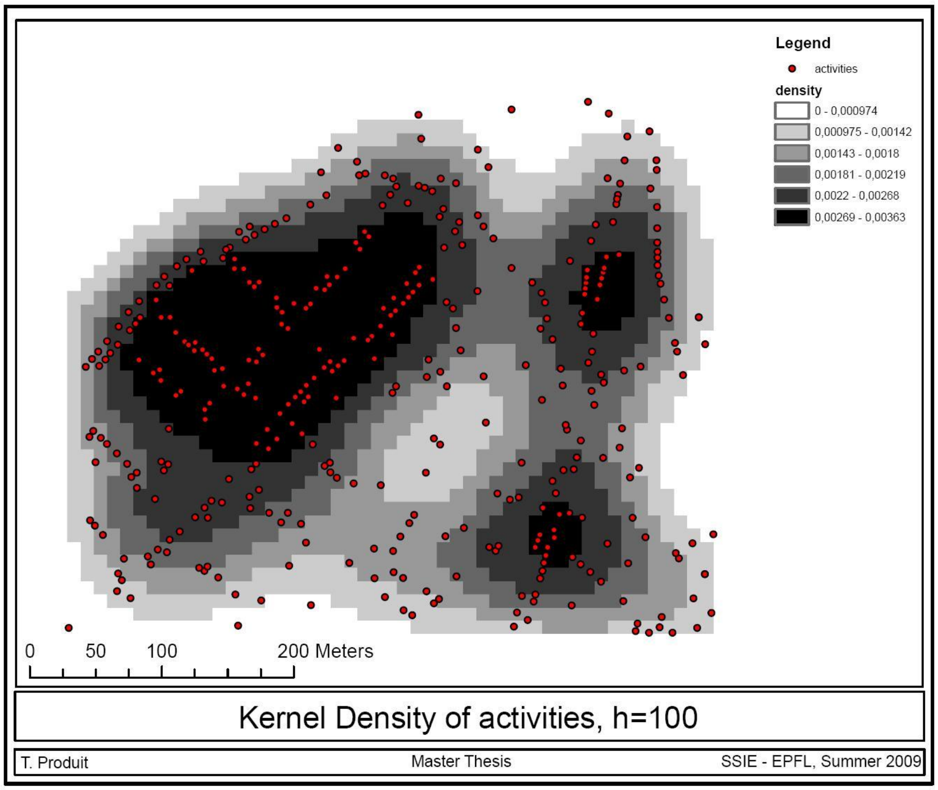


Figure 5.3: Kernel Density Estimation of the activities with a 100m bandwith

5.1.5 Network Density of activities

For the n*etwork density of activities*, the values are a kind of *kernel density* on the network. This means that the activities are weighted by their distance to the start point along the SPT. An activity near the start point has an high weight, and one far from the start point as a low weight (0 at the bandwith).
In the map on Fig 5.4, the network density of activities is given for a bandwith of 200m. Areas of high density of red dots (activities) show high values of density as well. Compared with the previous map, the high density areas follow the network to give an other perspective of the density. This view is more linked to the network.
In the map on Fig 5.5, page 52, the network density of activities is given for a bandwith of 100m. For this bandwith, results are not as intuitive as before. Probably, this bandwith is not adapted for this kind of network and activities distribution. It seems the number of activities and their weights have opposite effect and give granulary results.

After this glance on the density of activities, one can notice that the *kernel density* with a 100m bandwith is more related to the *network density* of 200m. The reason is related with the catchment area of the 200m SPT which is closer to a 100m disc than a 200m one.

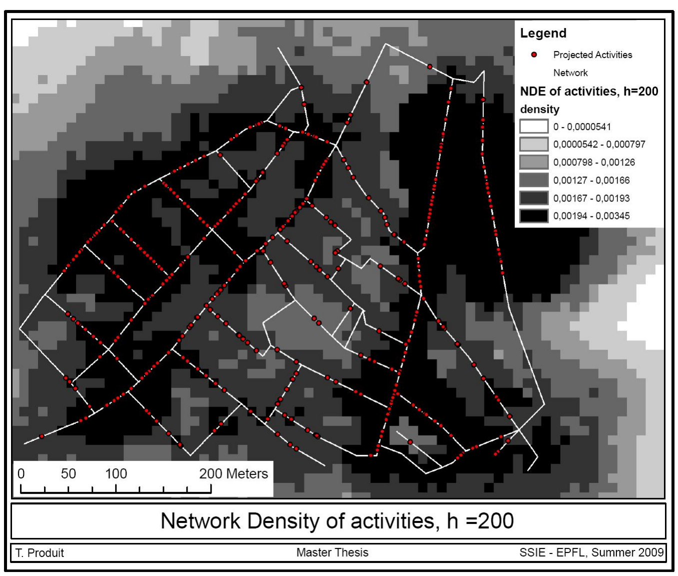


Figure 5.4: Network Density Estimation of the activities with a 200m bandwith

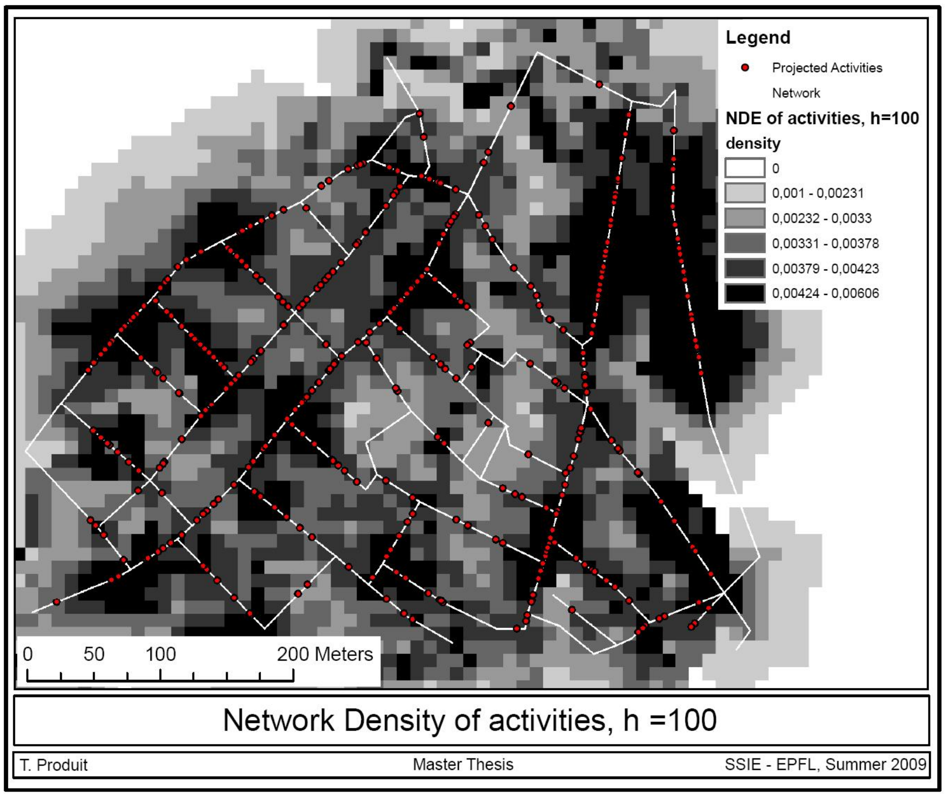


Figure 5.5: Network Density Estimation of the activities with a 100m bandwith

5.1.6 Kernel density of the edges

The map of the Fig 5.6 is the raster grid for the 100m *kernel density* of the edges, at present no value of population is used. This approach highlights the zone with many crossroads.

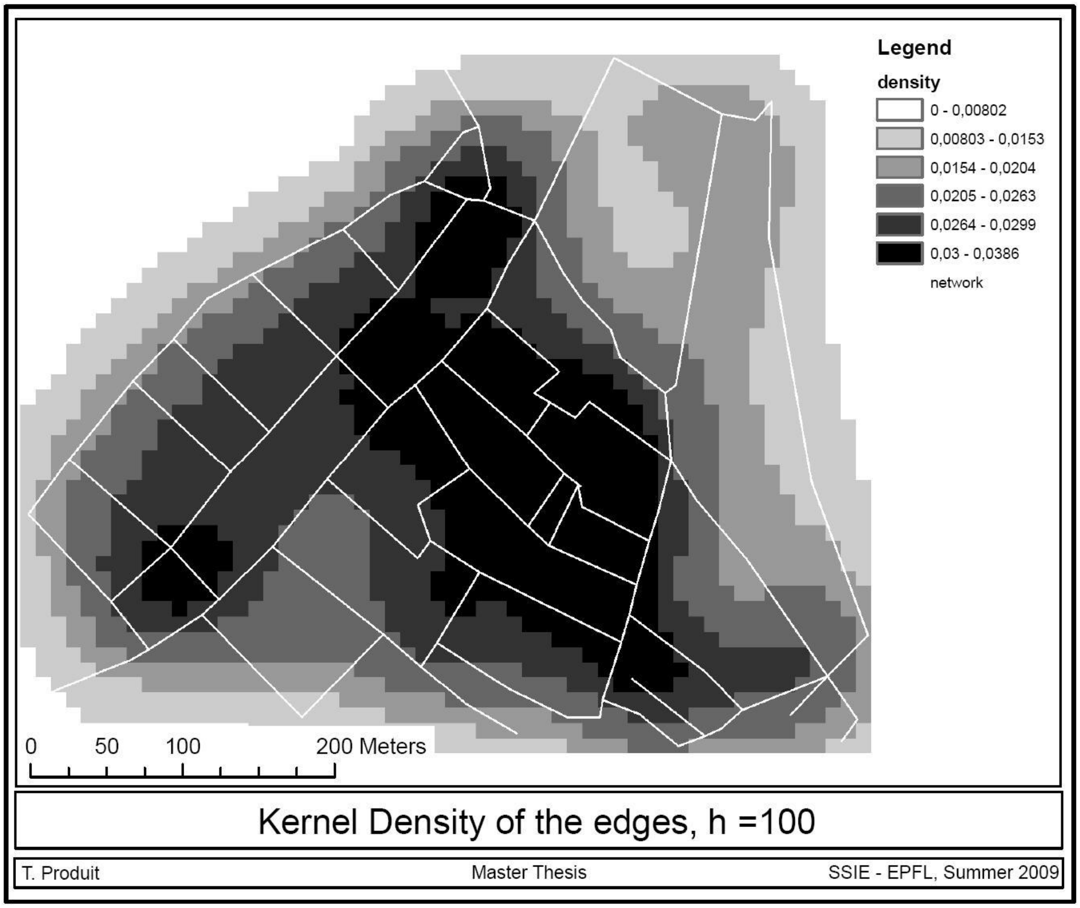


Figure 5.6: Kernel Density Estimation of the network with a 100m bandwith

5.1.7 Network density of edges

Before considering the *network density* of centrality, lets have a look at a more simple case, the *network density* of edges. A sort of *kernel density* is always used, nevertheless this time the focus is done at the edges: how many, how long and how far are they? *How far?* is the weighting, and *How long?* is a value of "population". This means that the long edges are more important than the short ones.
On the map of Fig 5.7, a bandwith of 100m is applied. Similary to the density of activities with 100m length, the scale do not fit well with this indicator. So it is hard to interpret the map. Lets have a look at the same indicator with a longer bandwith. The map Fig 5.8, page 55, represents the same indicator with a 200m bandwith. This scale seems to be more appropriate for this extract. Two SPTs of 200m are drawn to help understand the results. The yellow one has an high value of density because it is made from 17 edges. Most of them are near the start point and are quite long which explains this high value. In contrast, the red SPT has more edges (32) and a lot of them are short and far from the start point. In this way, these edges have a low weight and lead to a low value of edges density.

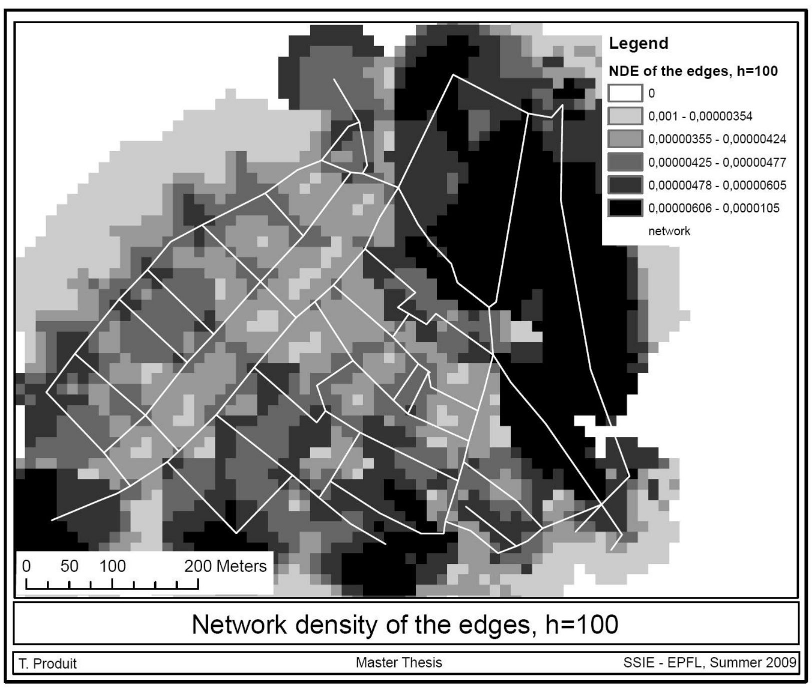


Figure 5.7: Network Density Estimation of the edges with a 100m bandwith

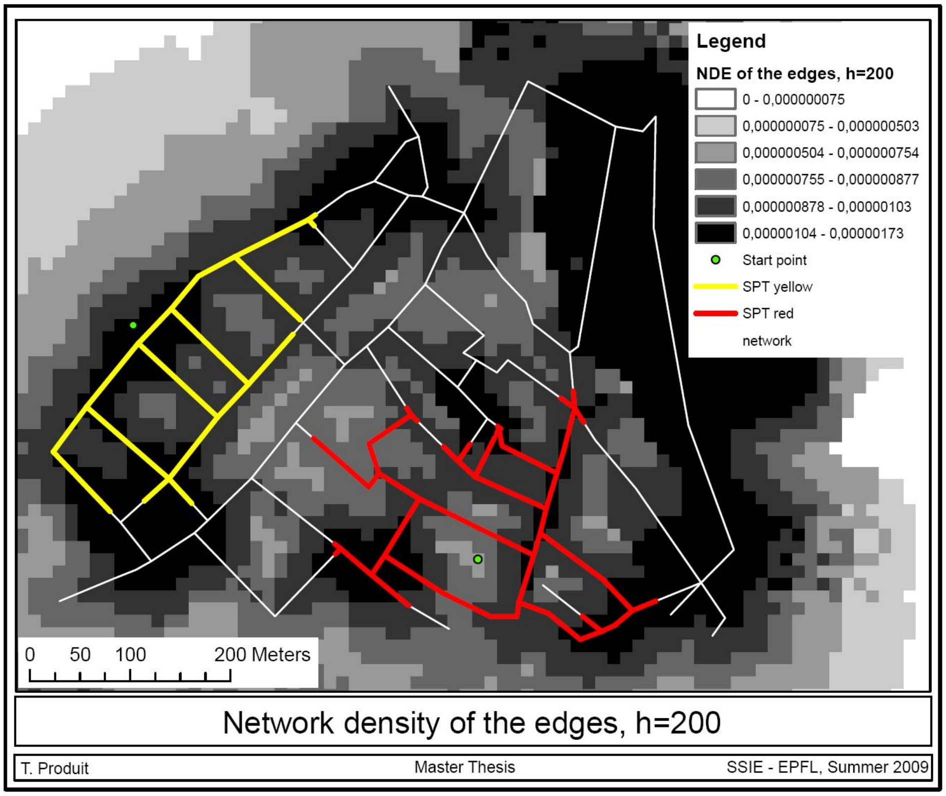


Figure 5.8: Network Density Estimation of the eges with a 200m bandwith

5.1.8 Kernel density of edges centralities

The centrality index used for this *kernel density estimation* is the global betweenness. Here (Fig 5.9), is the 100m *kernel density* of network centralities. The picture is really different from the picture of the KDE without values of population. This time, the patterns clearly follow the values of centralities.

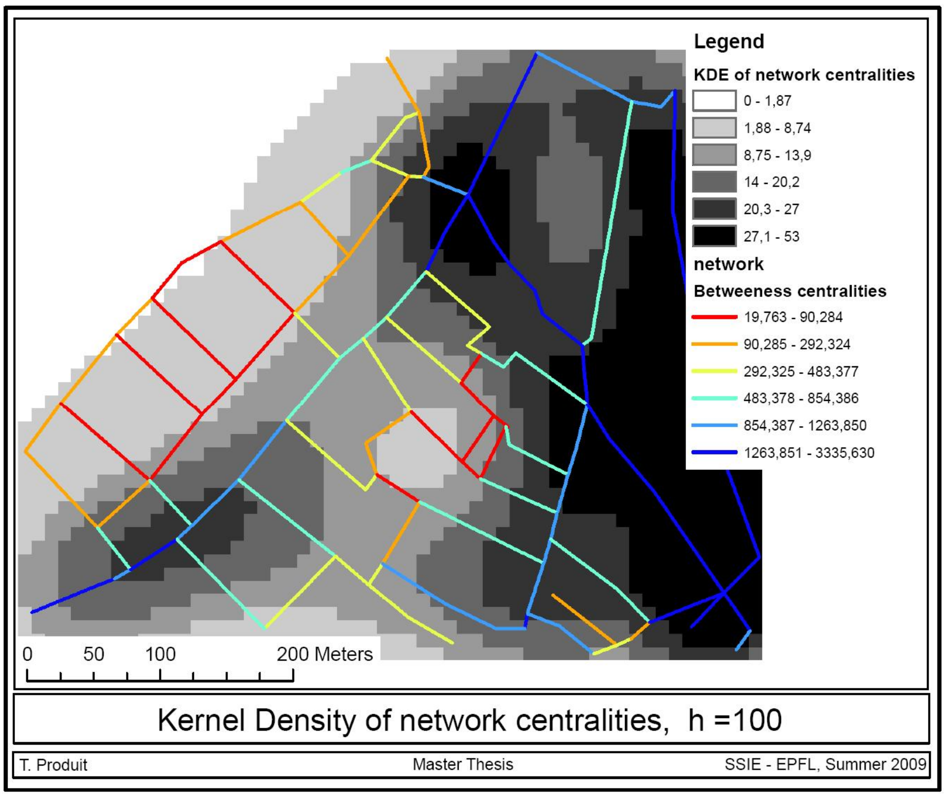


Figure 5.9: Kernel Density of network centralities with a 100m bandwith

5.1.9 Network density of edges centrality

Always using the global betweeness, the network density for a 200m bandwith is applied to this centrality. With the map in FIG 5.10, one can see that the results are close to the one of the KDE with a 100m bandwith. The patterns with a high value of density follow the high values of the network betweenness.

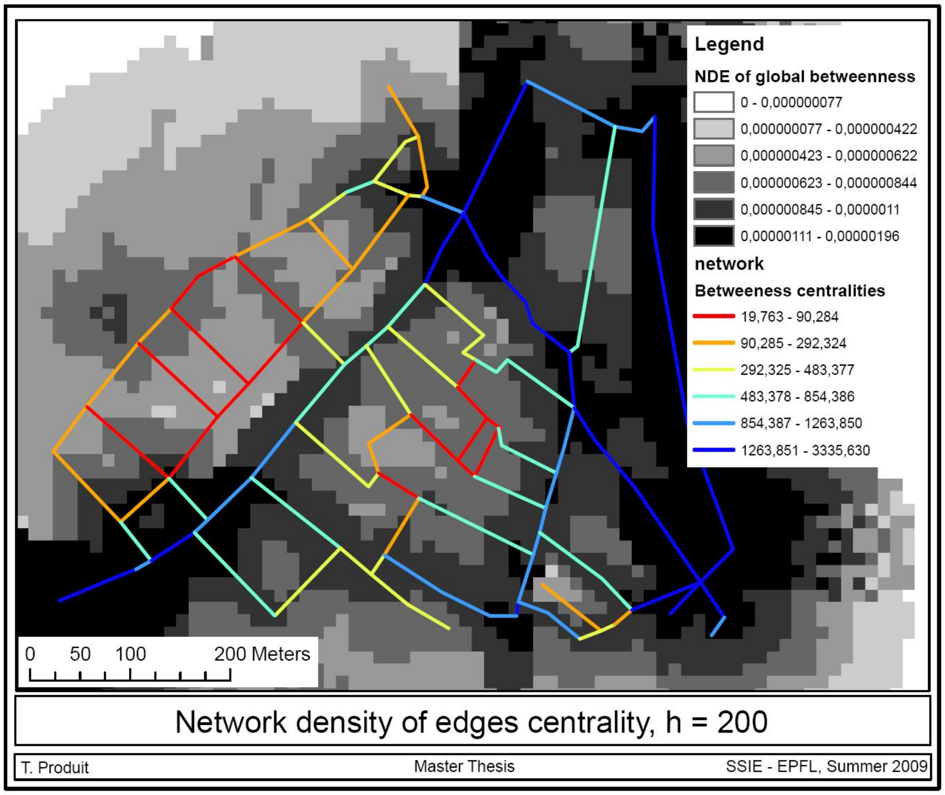


Figure 5.10: Network Density of network centralities with a 200m bandwith

5.1.10 Euclidean Diversity

For this index, only the shape of activities is used. These activities were sorted in six categories (*Detail store, Restaurant and reception, Computer science..., Wholesale business, Community activities and Others*). The diversity is computed from the activities belonging to a radius of 100m around each centroid of the raster. Shannon's diversity index was implemented in the algorithm.
First, one can see that the distribution of the activities follow the network, this fact justifies the network approach for events (here activities) in the urban space.

Shannon's indexes take into account the number of different activities and the ratio of specimens of one specie over all specimens. It explains well the low density in the bottom of the map FIG 5.11. Clearly, in this area the white and purple species are predominant and the other species have no many specimens. In contrast, in the top of the figure in the area of high diversity, you will find more species with more equal distribution of specimens.

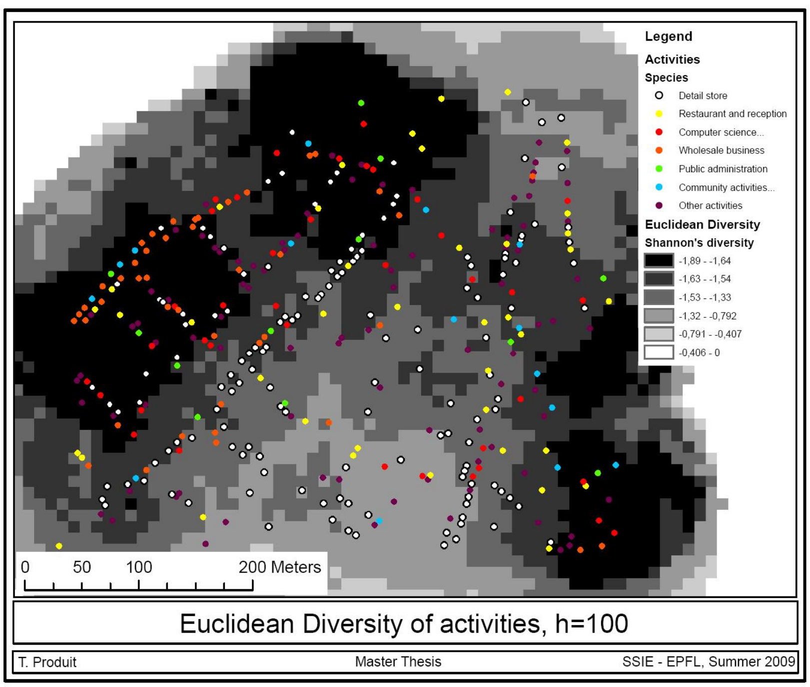


Figure 5.11: Euclidean diversity within a 100m bandwith

5.1.11 Network Diversity

In this second map of diversity, the activities are projected onto the network and the diversity is computed for the SPTs. In comparison with the *Euclidean Diversity*, one can see that the clouds of high diversity stay in the same locations. However, with this second approach the diversity fits better with the network and the clouds are arranged more accurately.

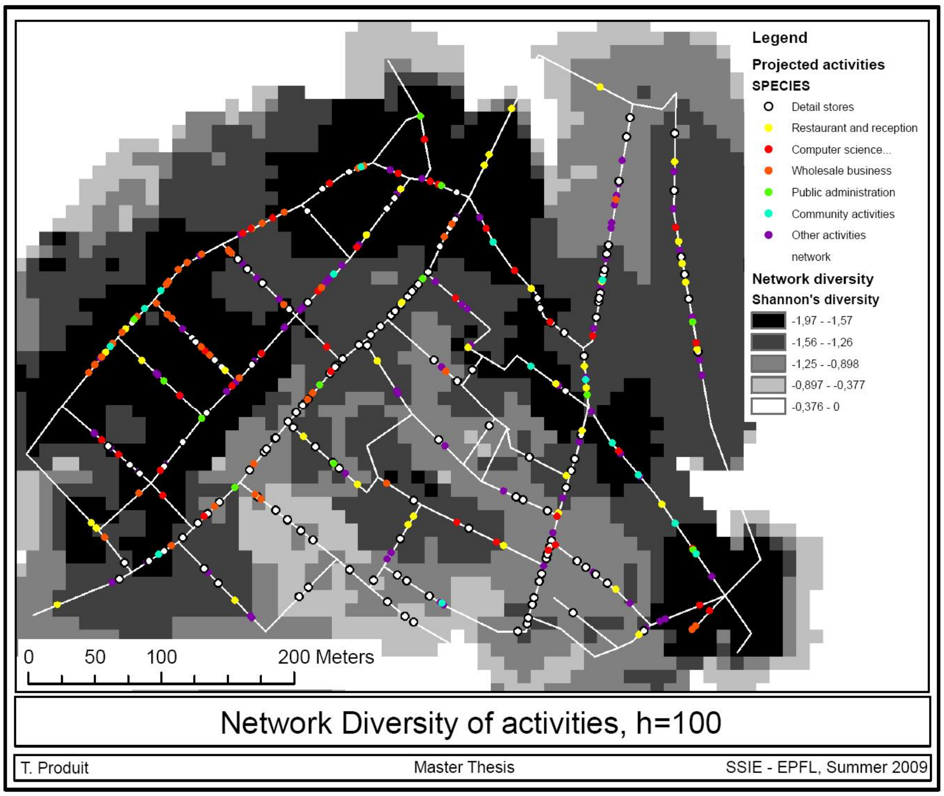


Figure 5.12: Network diversity with a 100m bandwith

5.2 Entire Barcelona

This section offers a quick overview of the Barcelona's context and presents the results of the entire Barcelona network. A first visual analysis is done between the maps to explain the results and try to determine the correlation results that can be expected.

5.2.1 Context

Some GIS facts

Once the algorithm provided good results on an extract of the network, it was applied to the entire Barcelona's network. This network is made of 11,222 edges for a total area of 92.65km^2. The *Agencia de Ecologia Urbana* give a database of 166,311 activities listed in 2002. This activities are sorted in seven categories:

1. Retail commerce;
2. Hotel, b&b, hostel, restaurant, pub...
3. IT, services to buisiness and people, research and development;
4. Gross commerce;
5. Public administration, education, health and social assitance;
6. Associational, recreational and sport activities;
7. Other activities (not related to public)

The presented calculation is for a 400m bandwith and here are some time considerations[1]:

- The projection of the 1,890,000 initial grid points takes 36 minutes. After this step, only 926,539 grid points remain, the others were too far from the network.
- The projection of the 166,311 activities takes 2 minutes.
- The computing of the euclidean diversity takes 3h30.
- The computing of the network densities and diversities takes 33 hours.

[1] For a description of the computer used, see section: A.2.1.

City of Barcelona

To have a good understanding of the results provided by the different calculations, it is necessary to condider the growth of Barcelona and what the city looks like today.

The first key is its geographical location. The city lies on a coastal plain and is constrained by the sea in the east, by hills in the west and rivers going from the hills to the sea. These constraints are the factors leading to the urban congestion and high residential densities that appear today in Barcelona.
The well-preserved medieval city (north part of the green shape, FIG 5.13, page 62) was built over the roman city and kept some characteristics of such cities, with alleyway and small squares. The south part of the green shape (*el Raval*) was the site for the industry during the early industrial revolution of the 19^{th} century. Barcelona is characterized by its tenement blocks influenced by the ideology of Karl MARX. In 1859, Idlefons CÉRDA planned these 520 blocks of *Eixample Garden City*. His goal was to provide equal acces to recreational space, shops, public transports and sunlight. Quickly, the garden city linked the old city with the outlying towns. Among them the town *Gracia* (blue shape) was finally annexed to Barcelona.

Today, downtown Barcelona can be divided into different areas. The grey shape is the *Eixample*; its south part is in the area of the *Pro-Eixample public open space initiative*, which aims to restore garden block interiors which have lost their original role during the unplanned growth of the last century. Since the 1992 Olympic Games, big investments are made in the inner city (green shape) to increase the living environment and favour social integration. The post-industrial area (yellow) redevelopped to become an high-technology incubator.
Some streets will be exhibited by the following map:

- **Yellow**: *Avinguda Diagonal*,
- **Orange**: *Gran Via Corts Catalanes*,
- **Dark blue**:*Avinguda de Gaudi*,
- **Light blue**: *Avinguda de Joseph Tarradelas*,
- **Green**: *Avinguda Meridiana*,
- **Red line**: Commercial area known as the shopping line of Barcelona. The principal street is *Passeig de Garcia*

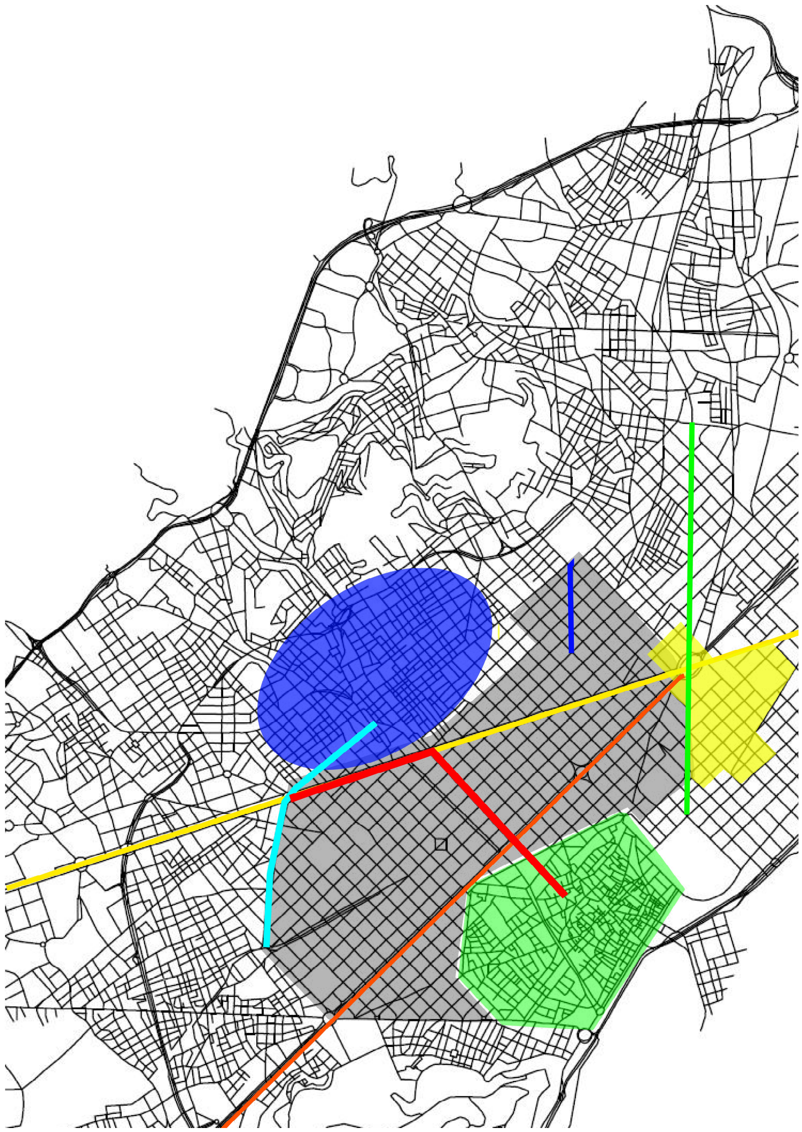


Figure 5.13: Yellow line: *Avinguda Diagonal*, **Orange line**: *Gran Via Corts Catalanes*, **Dark blue line**:*Avinguda de Gaudi*, **Light blue line**: *Avinguda de Joseph Tarradelas*, **Green line**: *Avinguda Meridiana*, **Red line**: Commercial area known as the shopping line of Barcelona, **Blue shape**: *Garcia*, **Grey shape**: *Eixample*, **Green shape**: inner city

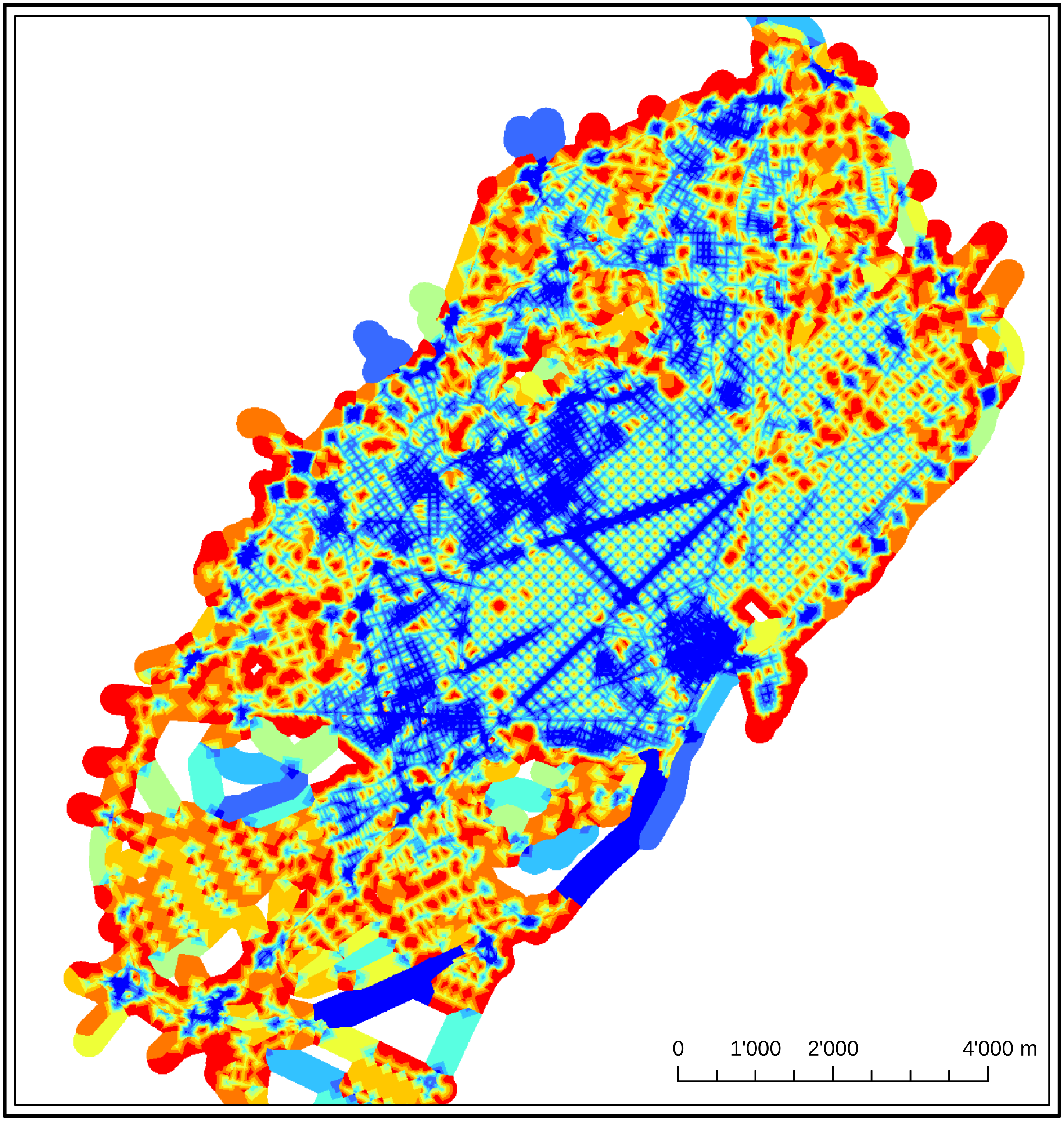
.

5.2.2 Total length of the SPT

On the two following maps, you can see the total length of the SPTs for the bandwiths 200m and 400m.
First, the two maps highlight the same locations of "high density". Nevertheless, the two scales are very different. The second one smooths the results in a bigger window, meaning that the original network better appears better in the first one.
Two kinds of geometry generate hot spots. The first one is the areas where the street density is high, like in the inner city or in *Garcia*. The second one are the areas where there are parallel edges very close, this geometry happens in the *Avinguda Diagonal* or in the *Passeig de Gracia*. Both these streets have two ways digitalized which increase the length of the SPTs [2].

[2] **Next maps:**

1. Total length of the SPTs, h=400.
2. Total length of the SPTs, h=200.



Legend

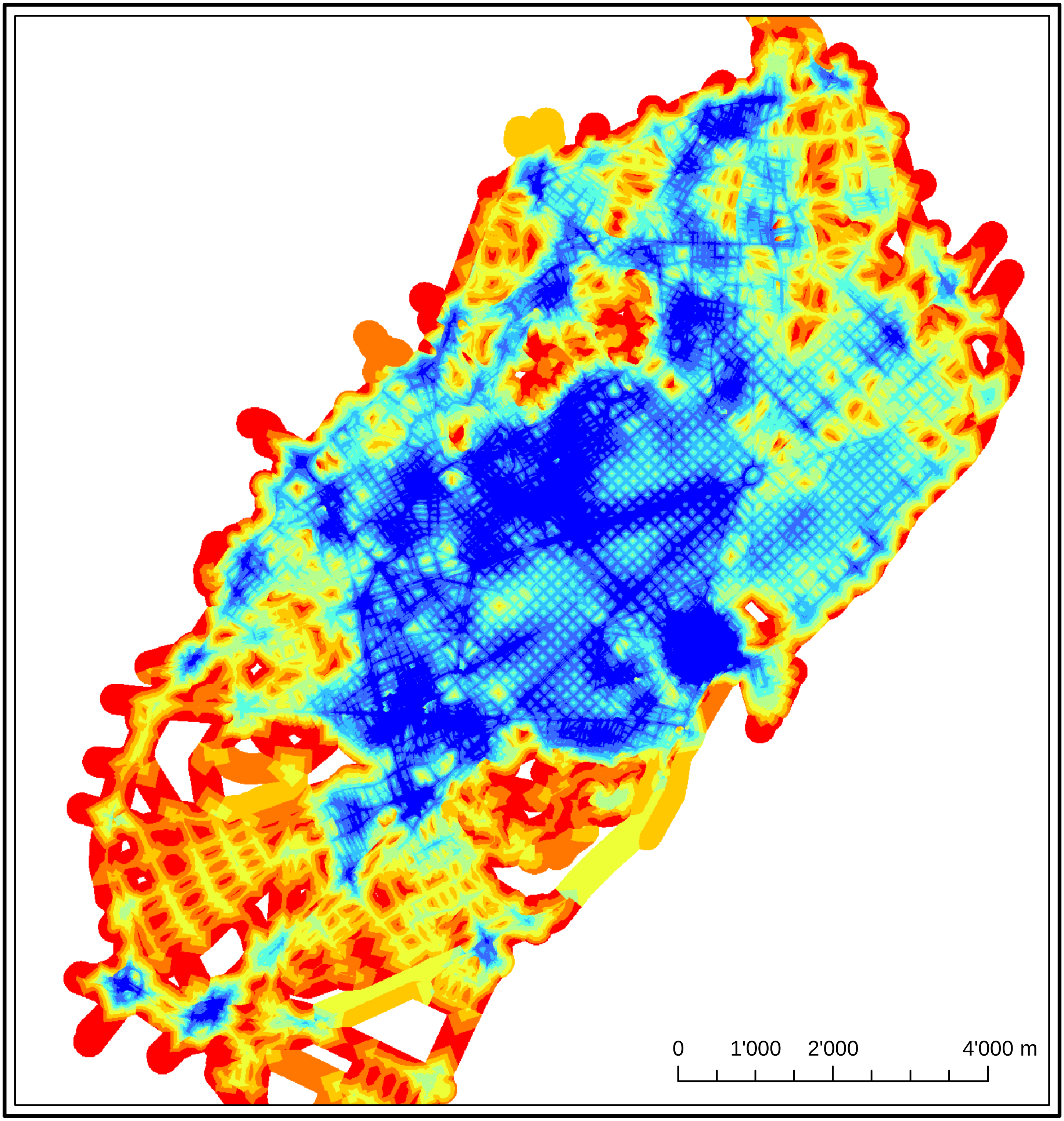
total length [m]

- 23 - 380
- 381 - 494
- 495 - 608
- 609 - 723
- 724 - 851
- 852 - 1'008
- 1'009 - 1'208
- 1'209 - 1'507
- 1'508 - 3'662

SPT total length, h = 200m

Comment:

T.Produit   Master Thesis   SSIE - LaSIG - EPFL



Legend

total length [m]

- 208 - 940
- 941 - 1'330
- 1'331 - 1'818
- 1'819 - 2'502
- 2'503 - 3'429
- 3'430 - 4'210
- 4'211 - 4'991
- 4'992 - 5'967
- 5'968 - 12'652

SPT total length, h = 400m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.3 Number of activities

To have a good visualization of the number of activities, this time the raster values were separated by equal intervals (Fig 5.14). The distribution of activities is logarithmic, there are a lot of raster cells with low values and a few with very high values. The equal interval avoid that high values disappear among the others at display. Moreover, high values have more interest for our work.
The *shopping line* of Barcelona is the blue patch which clearly appears in the two following maps. It goes from the inner city to the *Diagonal* through the *Passeig de Garcia* [3].

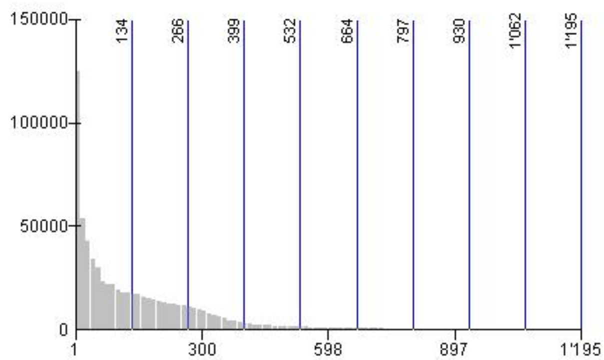
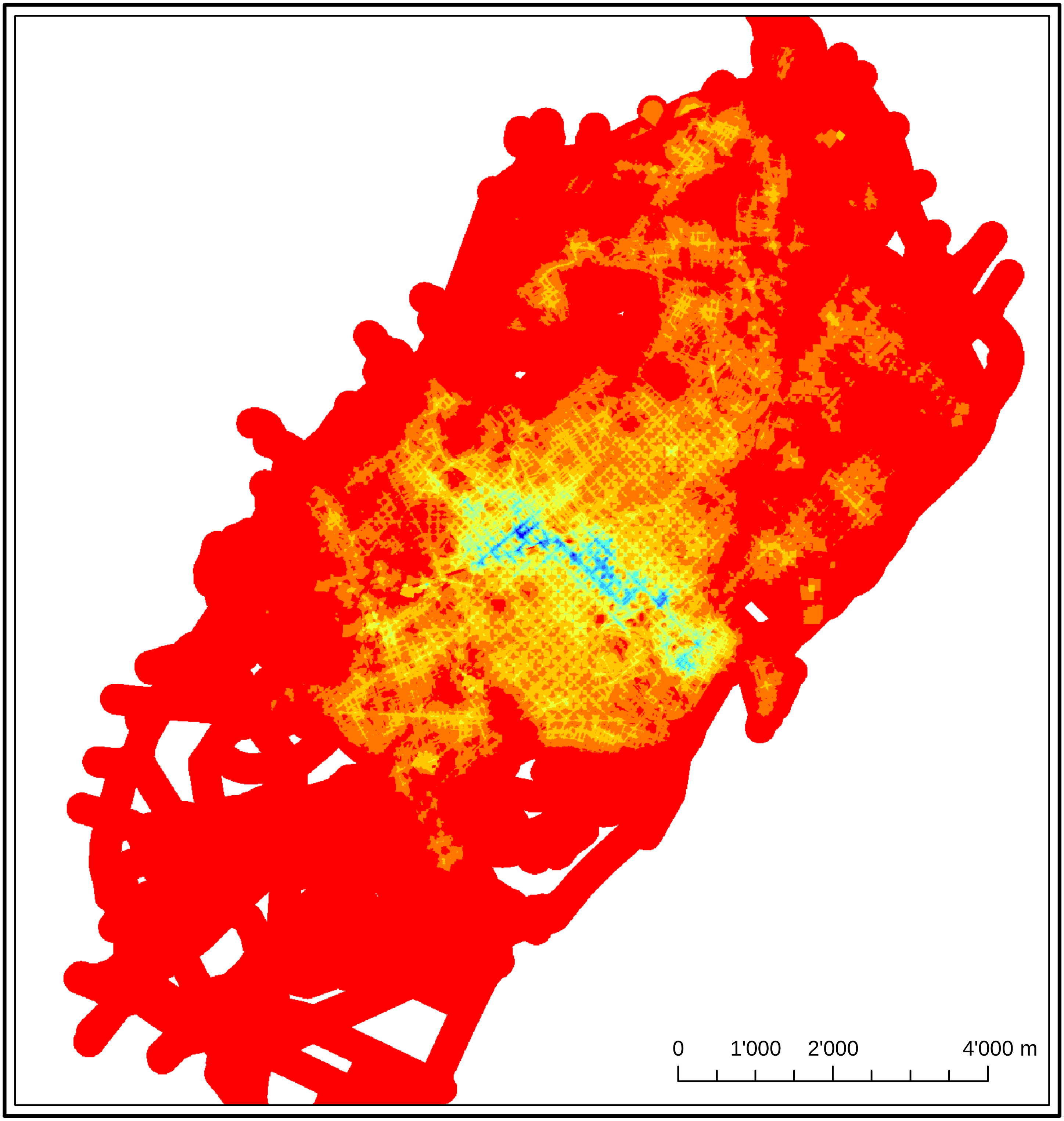


Figure 5.14: Distribution of the raster values and equal intervals for the number of activities in the 200m SPT

[3] **Next maps:**

1. Number of activities on the SPTs, h=400.
2. Number of activities on the SPTs, h=200.



Legend

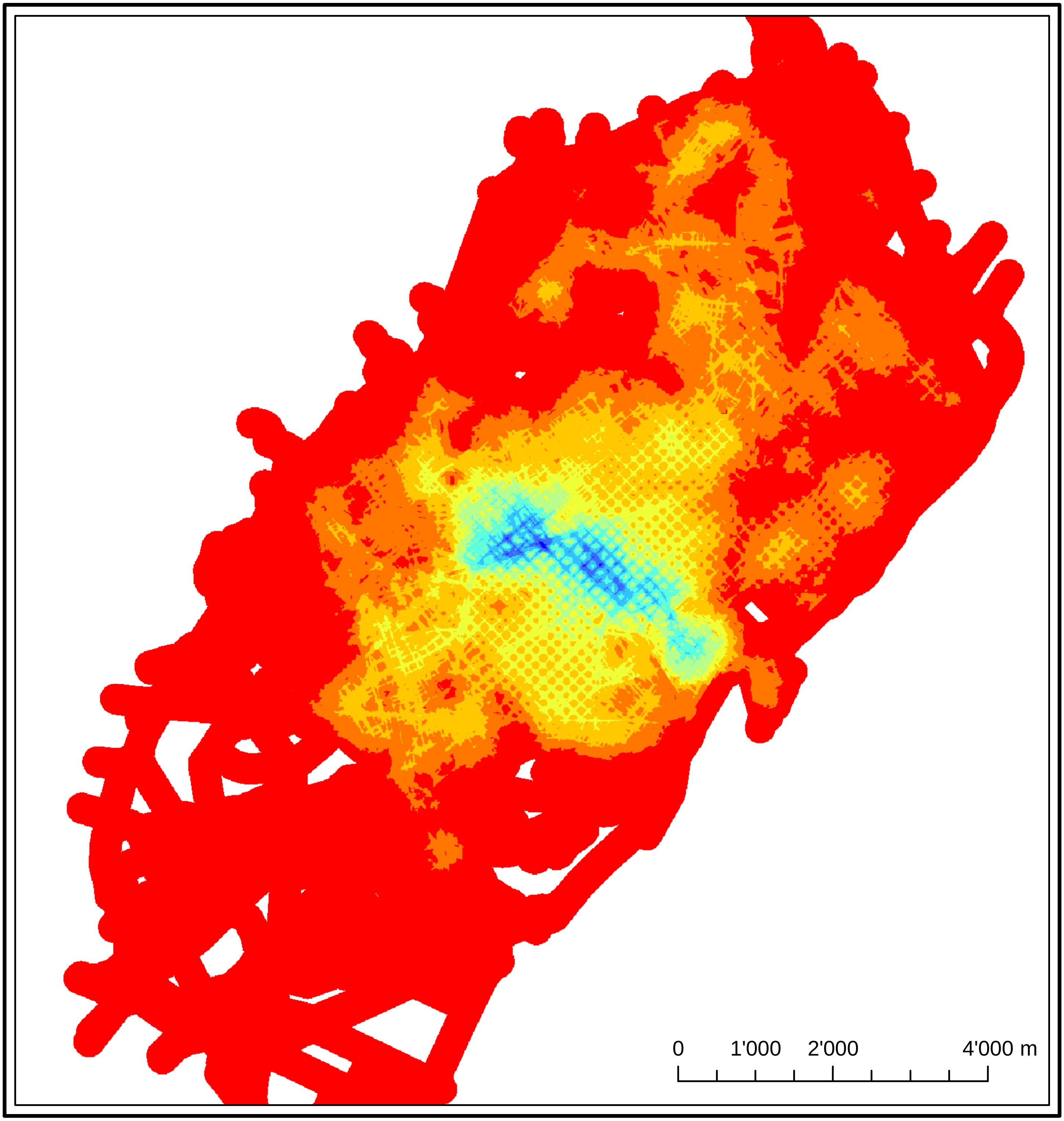
Number of activities

- 1 - 134
- 135 - 266
- 267 - 399
- 400 - 532
- 533 - 664
- 665 - 797
- 798 - 930
- 931 - 1'062
- 1'063 - 1'195

Number of activities, h = 200m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL



Legend

Number of activities

- 0 - 464
- 465 - 928
- 929 - 1'391
- 1'392 - 1'855
- 1'856 - 2'319
- 2'320 - 2'783
- 2'784 - 3'246
- 3'247 - 3'710
- 3'711 - 4'174

Number of activities, h = 400m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.4 Network density of activities

This index is very sensitive to the edge effect. So, in the following map dark red and dark blue shapes are mostly not representative. It is best to focus in the center of the city where the density of streets is sufficient to avoid such effects.
Remember that for this index the projected activities are used. In this way, the pixel near the network will have a higher density and the other one a lower density. Thus, the network geometry is brought to light. However, results are smoother in the inner city or everywhere around the *Eixample*, where the geometry create square patterns. The 400m bandwith gives a general information about density of activities and allows to discern patches of higher and lower density without big efforts. For example, the commercial line appears. The rendering may be better with a bigger bandwith. The results for the 400m bandwith are close to the 100m kernel density estimation. One more time, a scale appears between the network and the euclidean distance [4].

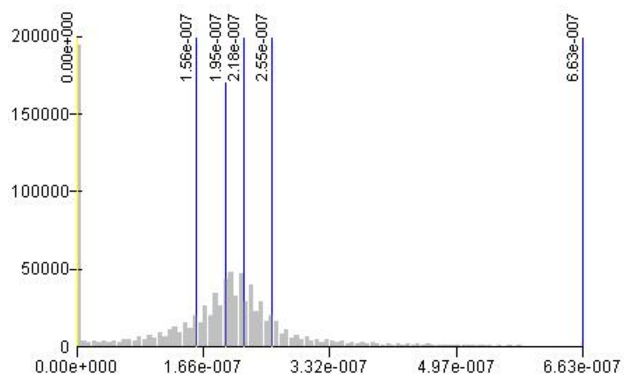
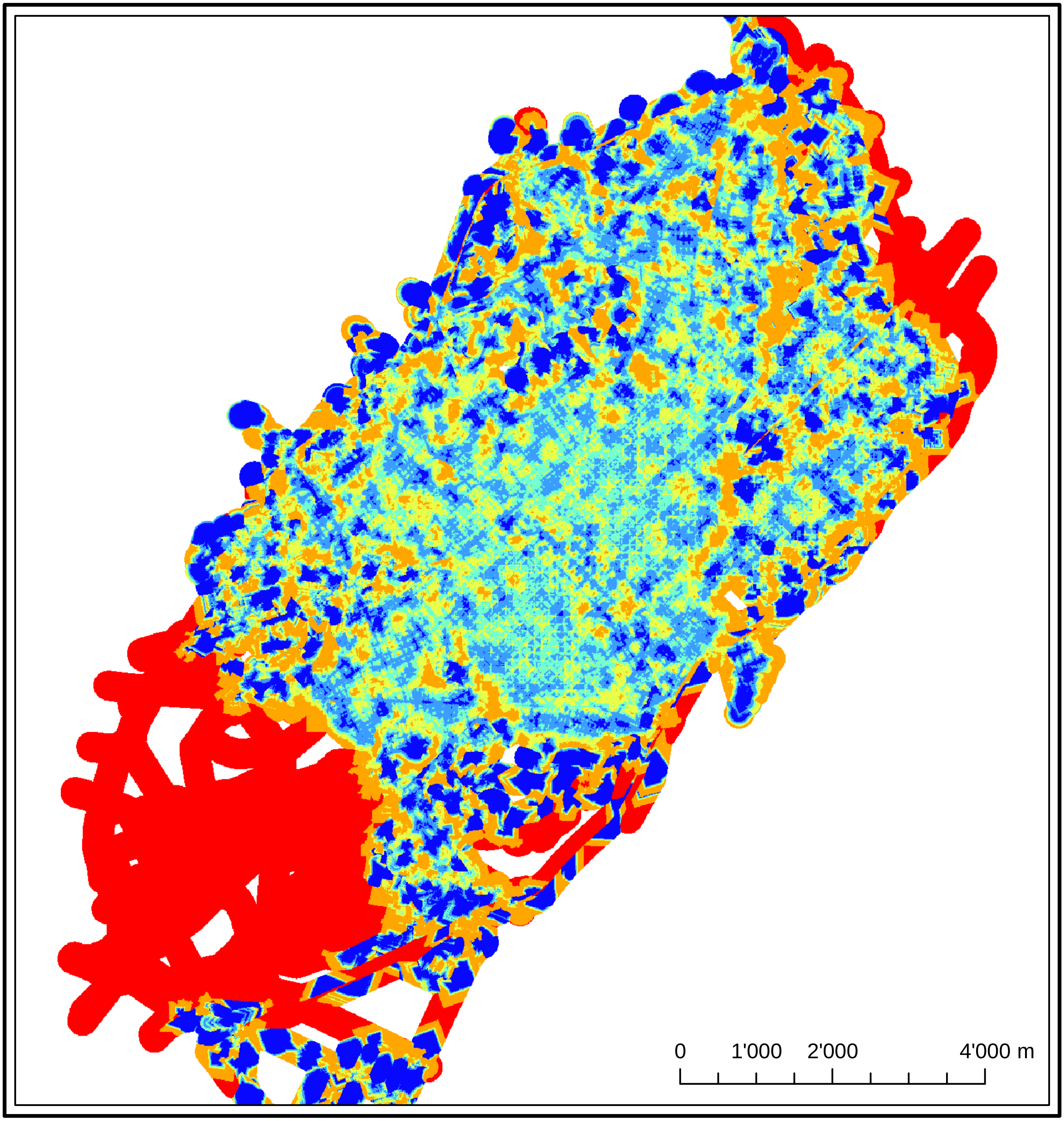


Figure 5.15: Distribution of the raster values and equal intervals for the number of activities in the 200m SPT

[4] **Next maps:**

1. NDE of activities with a 400m bandwith.
2. KDE of activities with a 100m bandwith.



Legend

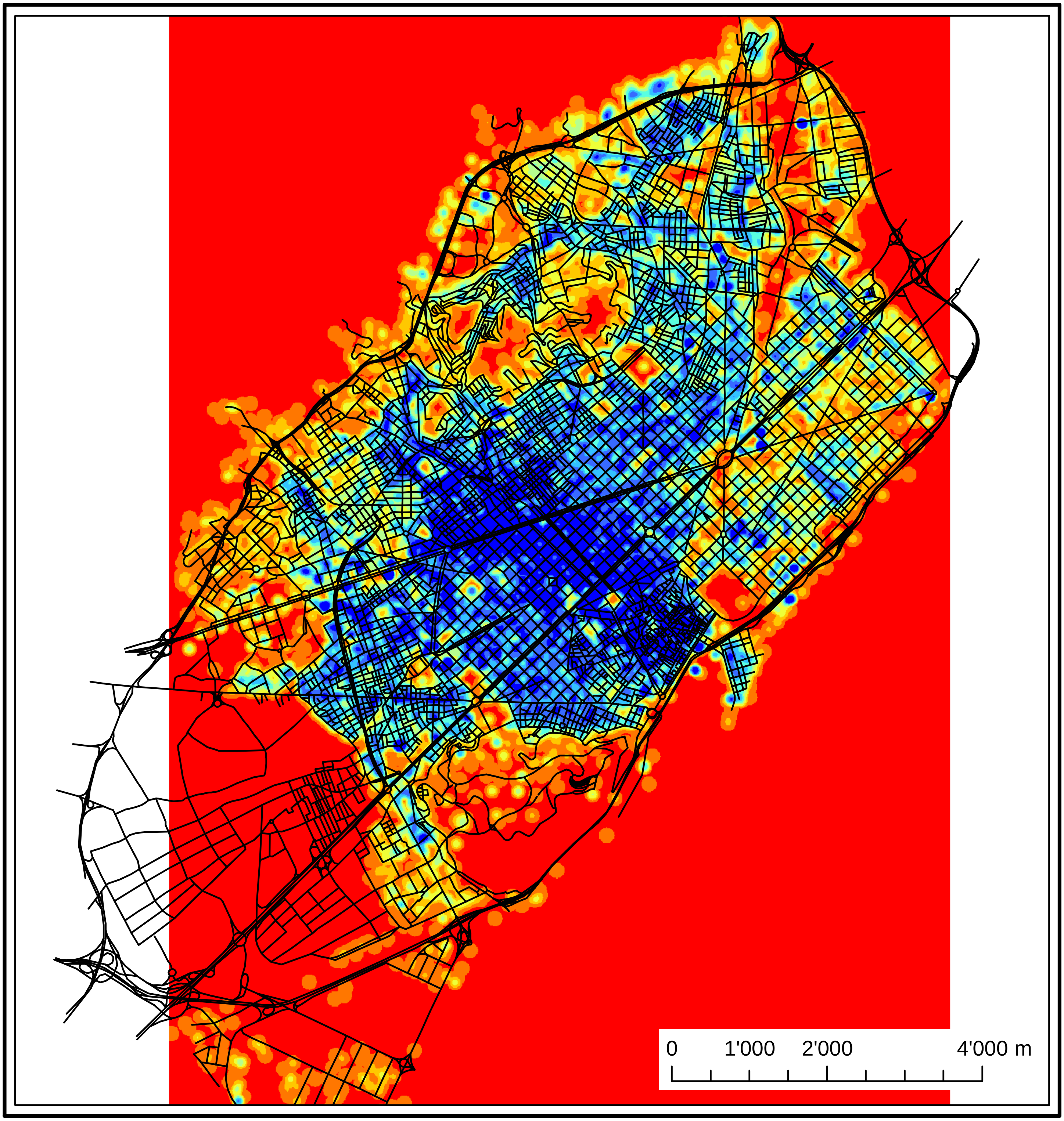
NDE of activities

- 0.00e+000
- 1.00e-002 - 1.56e-007
- 1.57e-007 - 1.95e-007
- 1.96e-007 - 2.18e-007
- 2.19e-007 - 2.55e-007
- 2.56e-007 - 6.63e-007

Network density of activities, h = 400m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL



Legend

network

density of activities

- 0
- 0.001 - 0.000312
- 0.000313 - 0.000833
- 0.000834 - 0.00156
- 0.00157 - 0.0025
- 0.00251 - 0.00343
- 0.00344 - 0.00447
- 0.00448 - 0.00593
- 0.00594 - 0.0265

Kernel density estimation of the activities, h = 100

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.5 Network density of edges betweenness centrality

Except in areas where density of the edges is low and so egde effects are high, the *network density* of betweenness allows to have a good visualization of this centrality. The results are strongly network-dependent. On the map with the 400m bandwith, the raster exhibits the locations with high value of centrality, but the network approach allows to make out the network as well. This index give values close to the reality of the city.

The high values of centrality for most of them match well with the big streets of Barcelona. The *Avinguda Diagonal* and the *Gran Via de les Corts Catalanes* are such streets. For some others, like the *Avinguda de Gaudi* the high value of centralities fits with little street. For this little street, the geometry of the network plays a big role: it cuts the regular geometry of the *Eixample* and is a shortcut. An other interesting shape is the area near the *Avinguda Meridiana*, the centrality is shifted to little parallel streets. Nervertheless, the direction stays the same

The results are close to the one of the standard KDE with a 100m bandwith, the same streets are highlighted [5].

[5] **Next maps:**

1. Global betweenness of the network.
2. NDE of global betweenness, h = 400.
3. KDE of global betweenness, h = 100.

The distribution of the global betweenness values and NDE of global betweenness are presented in FIG 5.17 and FIG 5.16. Keep in mind that the distribution of the global betweenness is logarithmic.

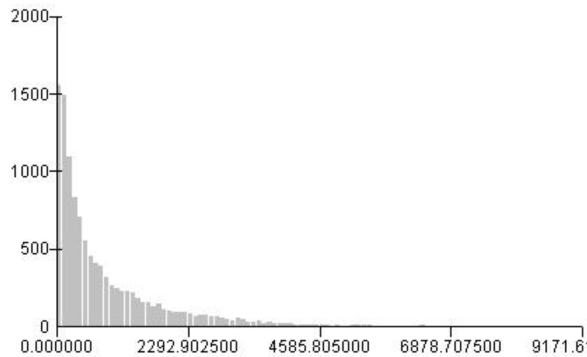


Figure 5.16: Distribution of the edges values for the global betweenness

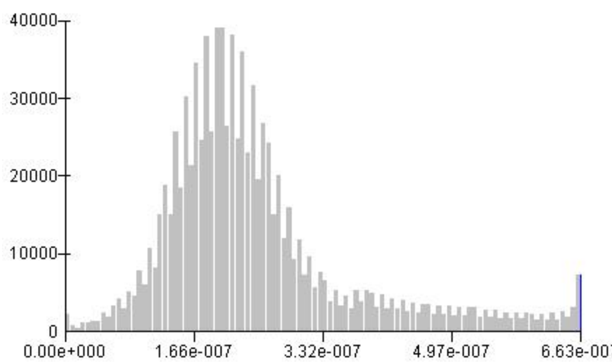
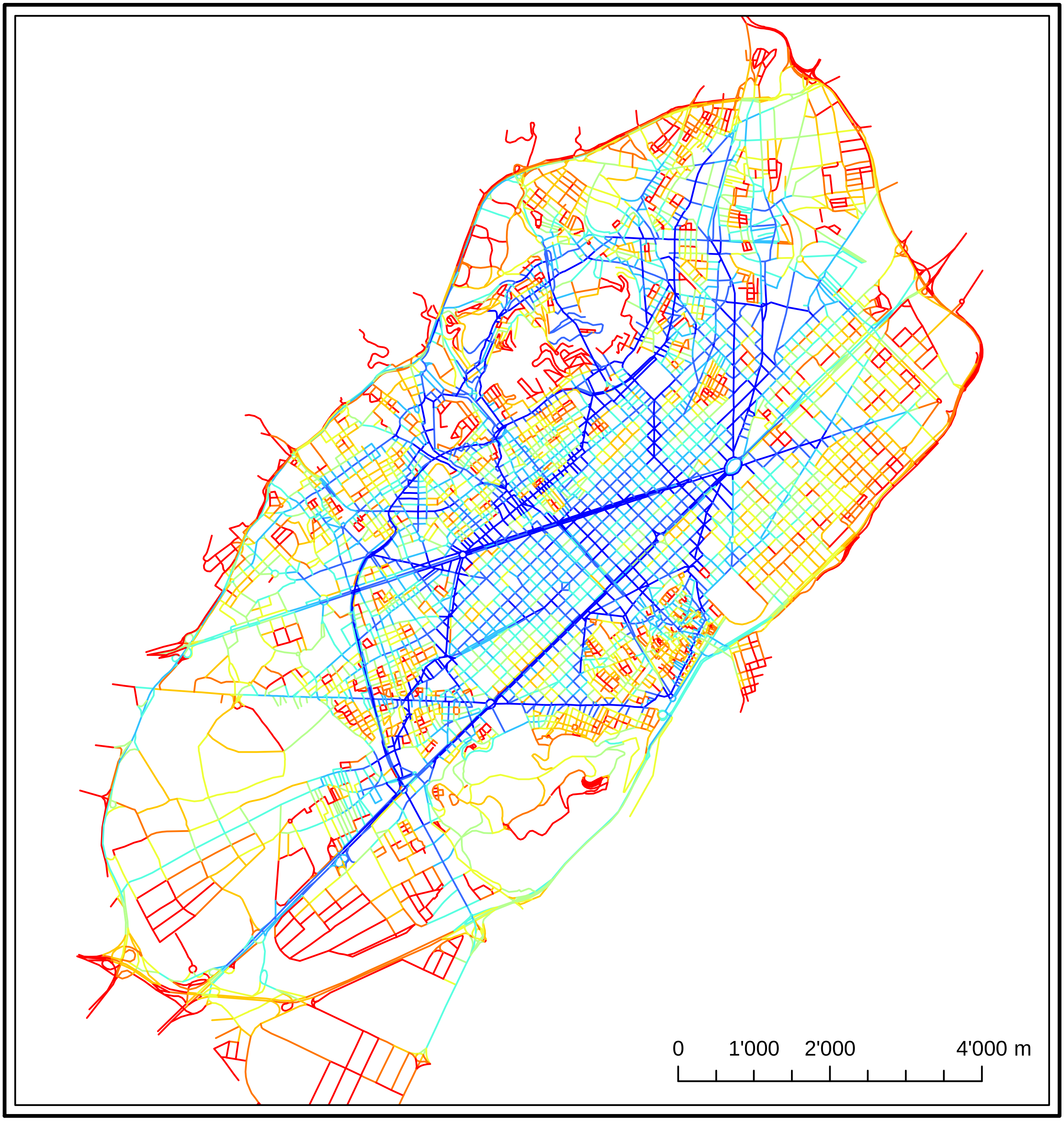


Figure 5.17: Distribution of the raster values for the network density of global betweenness with a 200m SPT



Legend

network

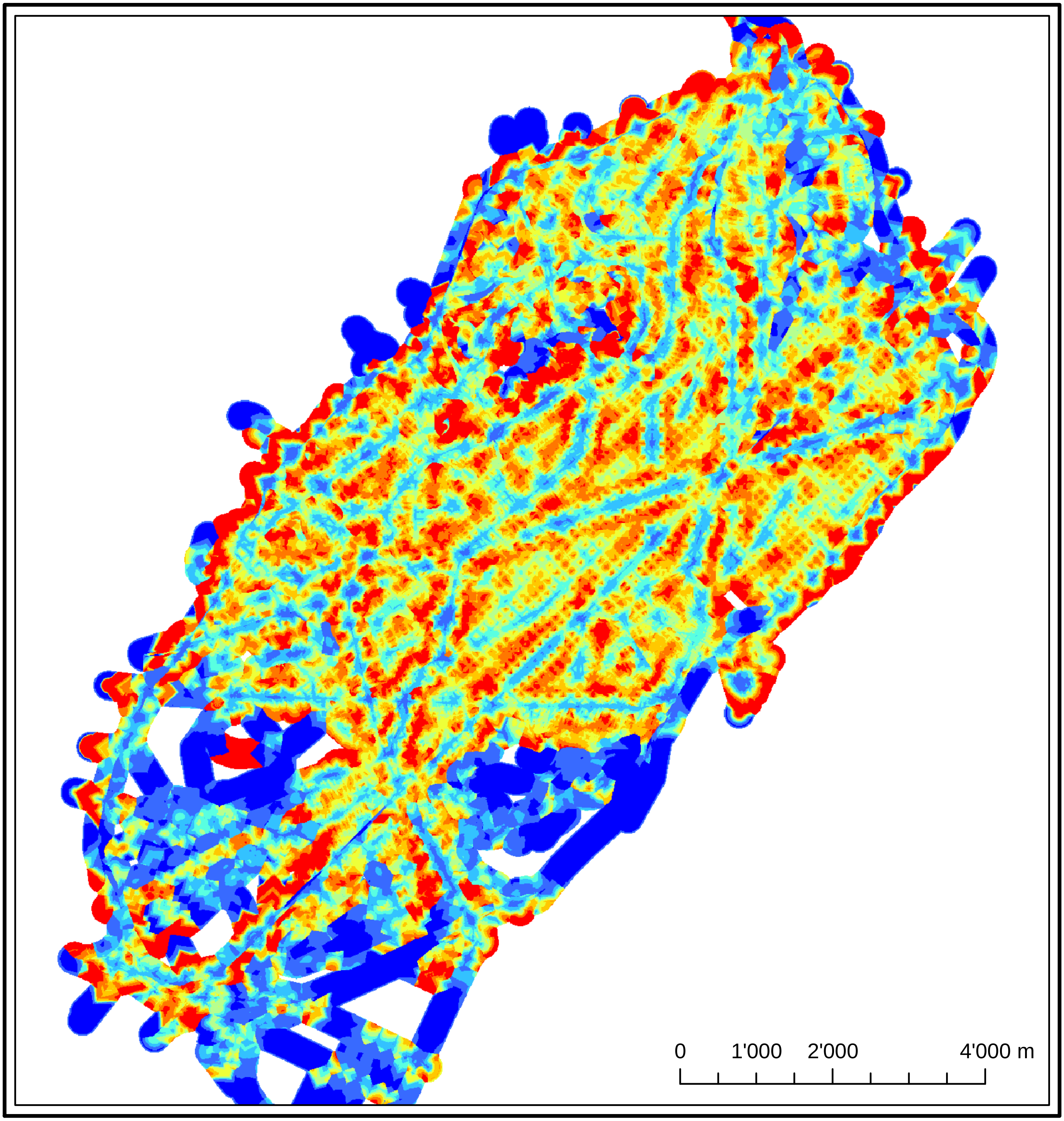
betweenness centrality

- 0.000000000 - 76.1
- 76.2 - 148
- 149 - 240
- 241 - 370
- 371 - 553
- 554 - 828
- 829 - 1280
- 1290 - 2060
- 2070 - 9170

Global betweenness centrality

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL



Legend

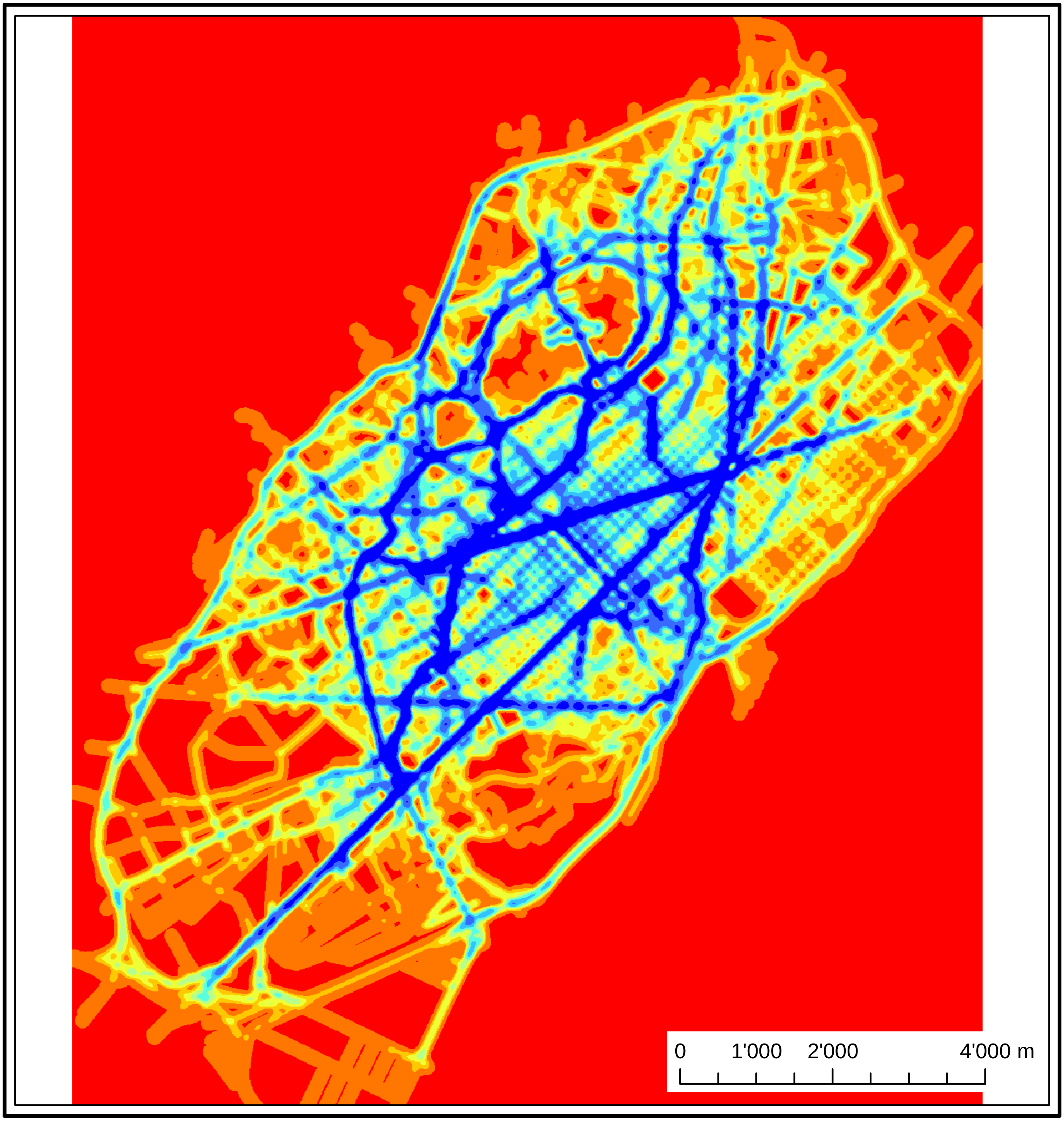
NDE of global betweenness

0.00e+000 - 1.35e-007
1.36e-007 - 1.64e-007
1.65e-007 - 1.87e-007
1.88e-007 - 2.08e-007
2.09e-007 - 2.31e-007
2.32e-007 - 2.57e-007
2.58e-007 - 2.99e-007
3.00e-007 - 4.21e-007
4.22e-007 - 6.63e-007

Network density of global betweenness, h = 400m

Comment:

T.Produit   Master Thesis   SSIE - LaSIG - EPFL



Legend

KDE of global betweenness

- 0
- 0.001 - 2.03
- 2.04 - 4.05
- 4.06 - 7.09
- 7.1 - 11.1
- 11.2 - 16.2
- 16.3 - 23.3
- 23.4 - 38.5
- 38.6 - 258

Kernel density of global betweenness, h = 100m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.6 Euclidean diversity

This index is able to make out some distinctive areas of Barcelona. The easiest pattern to explain is the *Passeig de Garcia* which is the commercial street linking *Garcia* with the inner city. This boulevard is known as the shopping line of Barcelona. There, the diversity is low, because the predominant specie is the *Servives* and *Detail stores* familly. *El Carmel* in the north is a rural neighbourhood, which is essentially residential, so this area is less attractive for some activities. One can make out some circles on the map, these circles highligth the location of an important building or activity. For example, one of them is a theatre, where a lot of music, dance or theater associations are located. In this way, a very specialized building like a theater or a public administration in which a lot of specimens of the same specie are located, can have such a visible diversity impact.
In the two following maps, the colors are representing the same values for a better comparison. In general, the diversity increase with the bandwith. This means that the city is specialized at low bandwith. Nevertheless, location of high diversity at low bandwith stays in the same values or slightly decreases [6].

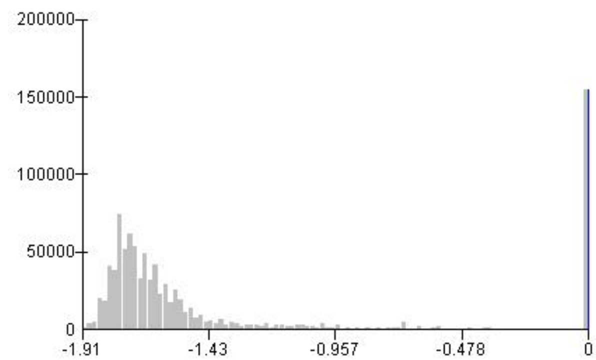
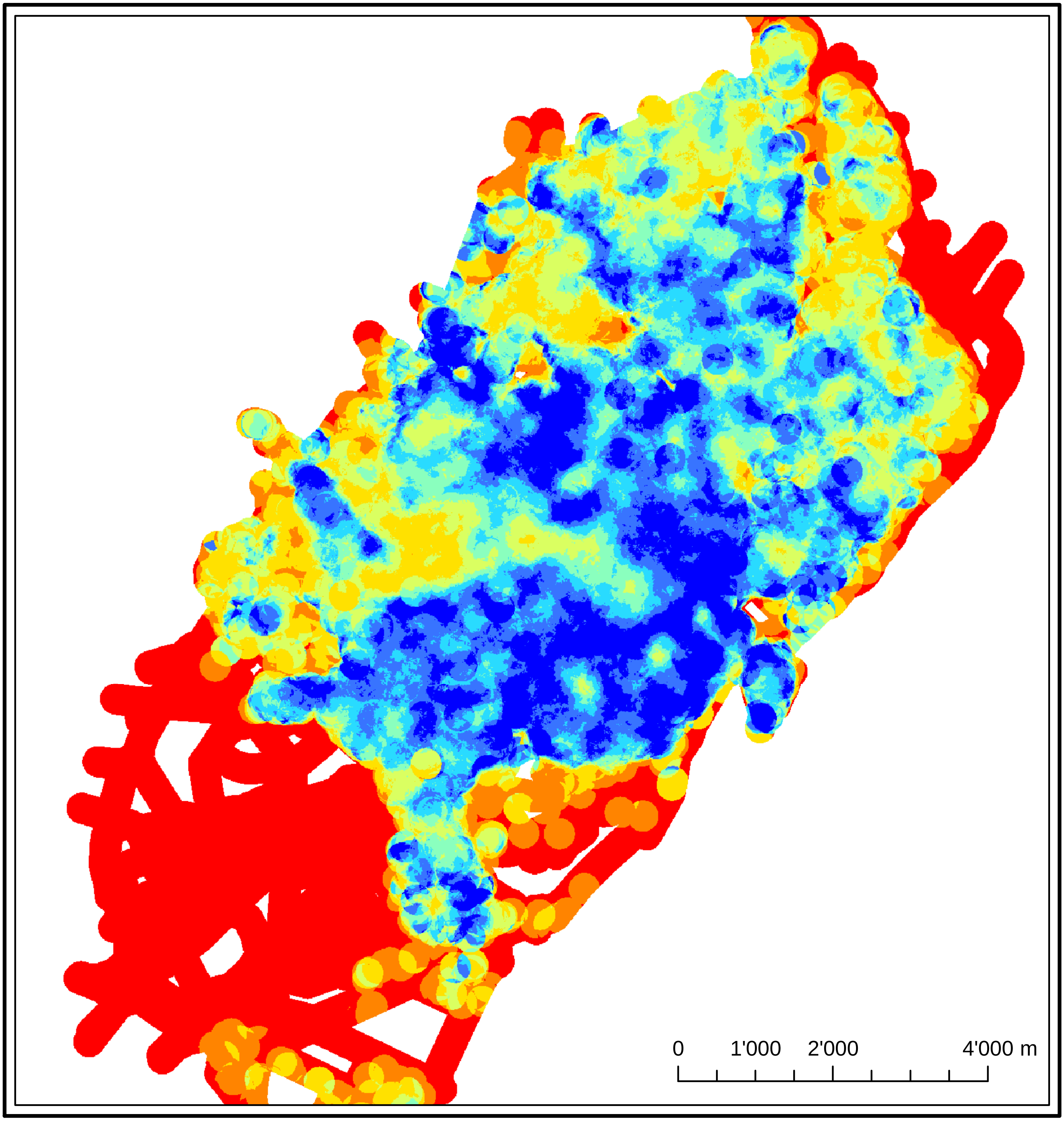


Figure 5.18: Distribution of the raster values for the euclidean entropy of activities on the 400m SPT

[6] **Next maps:**

1. Euclidean diversity, h = 200.
2. Euclidean diversity, h = 400.



Legend

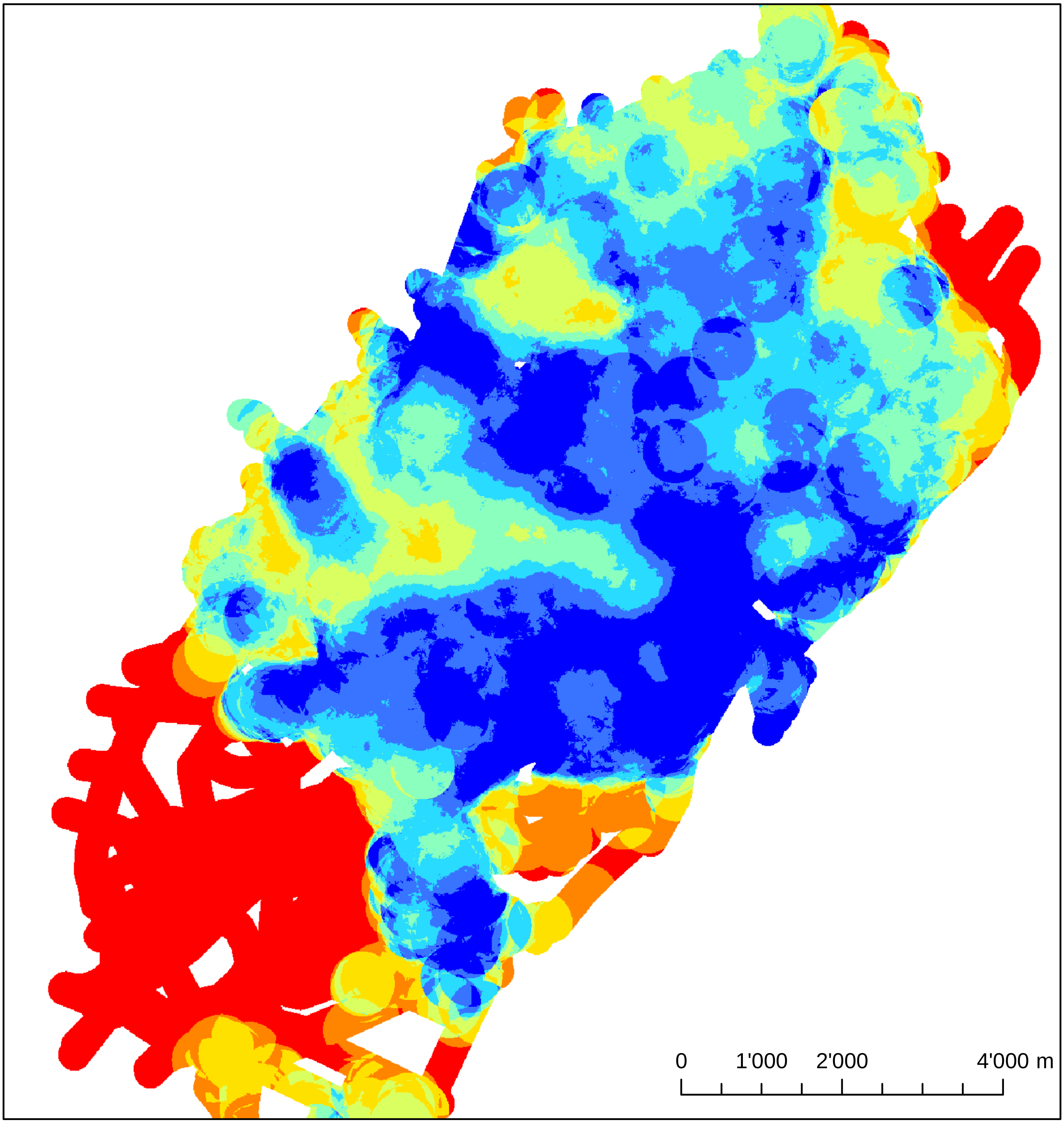
Entropy

- -1.93 - -1.77
- -1.76 - -1.72
- -1.71 - -1.66
- -1.65 - -1.59
- -1.58 - -1.48
- -1.47 - -1.16
- -1.15 - -0.128
- -0.127 - 0

Euclidean entropy, h = 200m

Comment:

T.Produit | Master Thesis | SSIE - LaSIG - EPFL



Legend

Entropy

Euclidean entropy, h = 400m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.7 Network diversity

The network diversity shows the same locations than the euclidean diversity. However, this index is less sensitive or looks less sensitive to unique locations. The impact of such particular locations, which lead to visible circle on the previous maps, are less important with this index.
The network diversities are always lower for a same bandwith than the euclidean one. This is always because of the scale between the catchment area of the SPT and the circle for the same bandwith.

To go further in the interpretation two other parameters are calculated. First, the predominant specie, this value simply represent the biggest specie in the sample. Then, the next paramater is the ratio of this specie.
Thus, a better interpretation is possible. For example, near the commercial line the specie *Professional or IT activities, services...* is the most important, it can reach 60%. West the inner city, there is an other patch of lower diversity, there, the *Detail sores* represents between 40 and 60% of the activities[7].

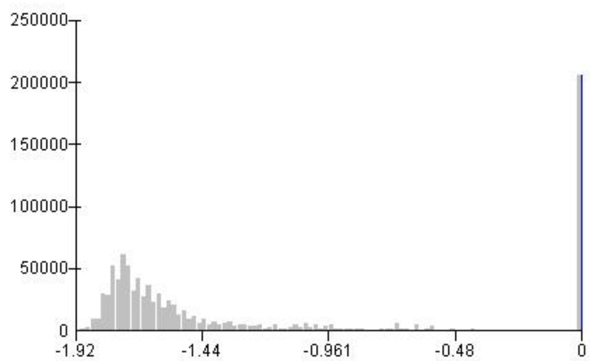
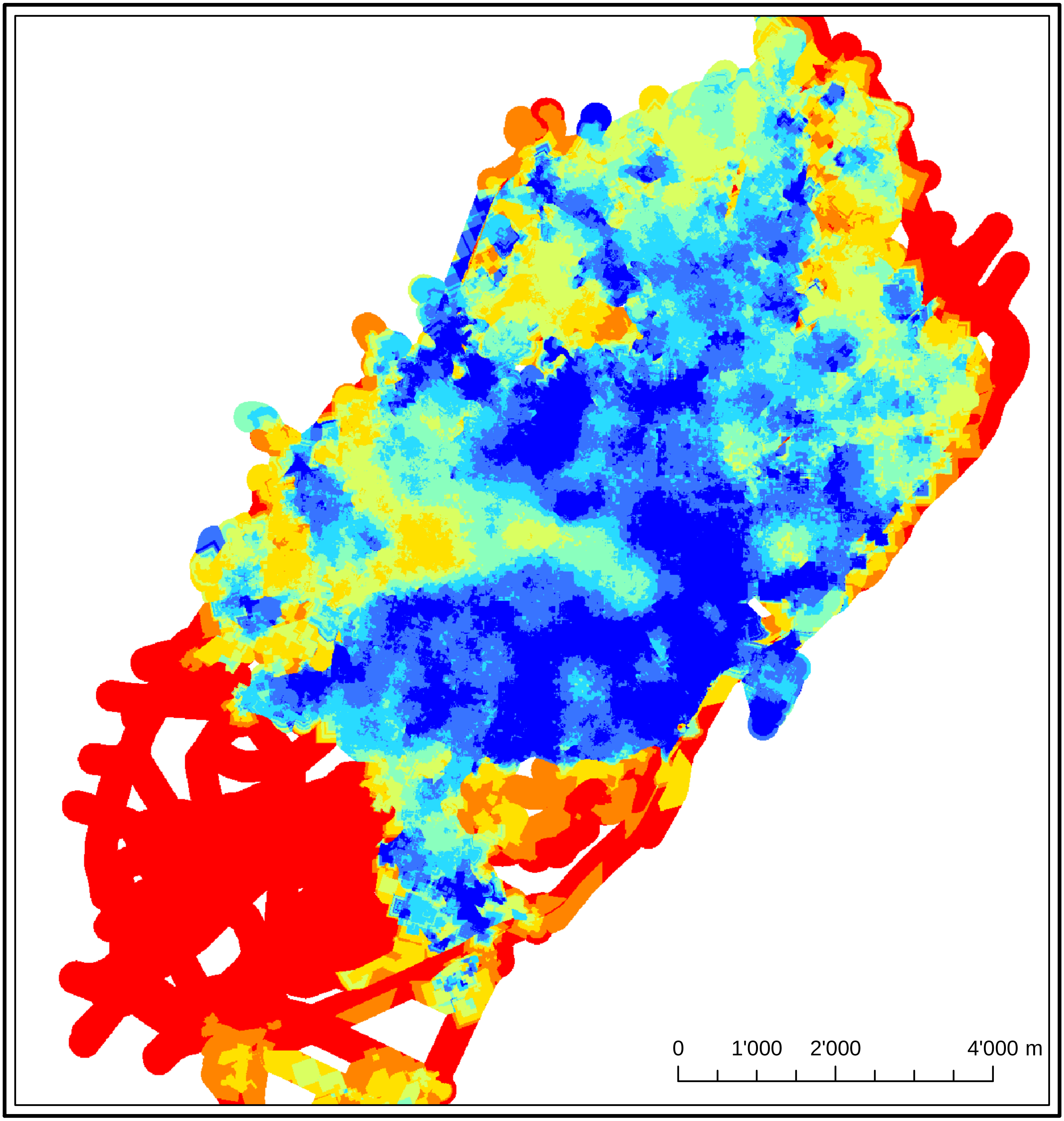


Figure 5.19: Distribution of the raster values for the network diversity of activities with a 400m SPT

[7] **Next maps:**

1. Network diversity, h = 400.
2. Network diversity: predominant specie, h = 400.
3. Network diversity: predominant specie ratio, h = 400.



Legend

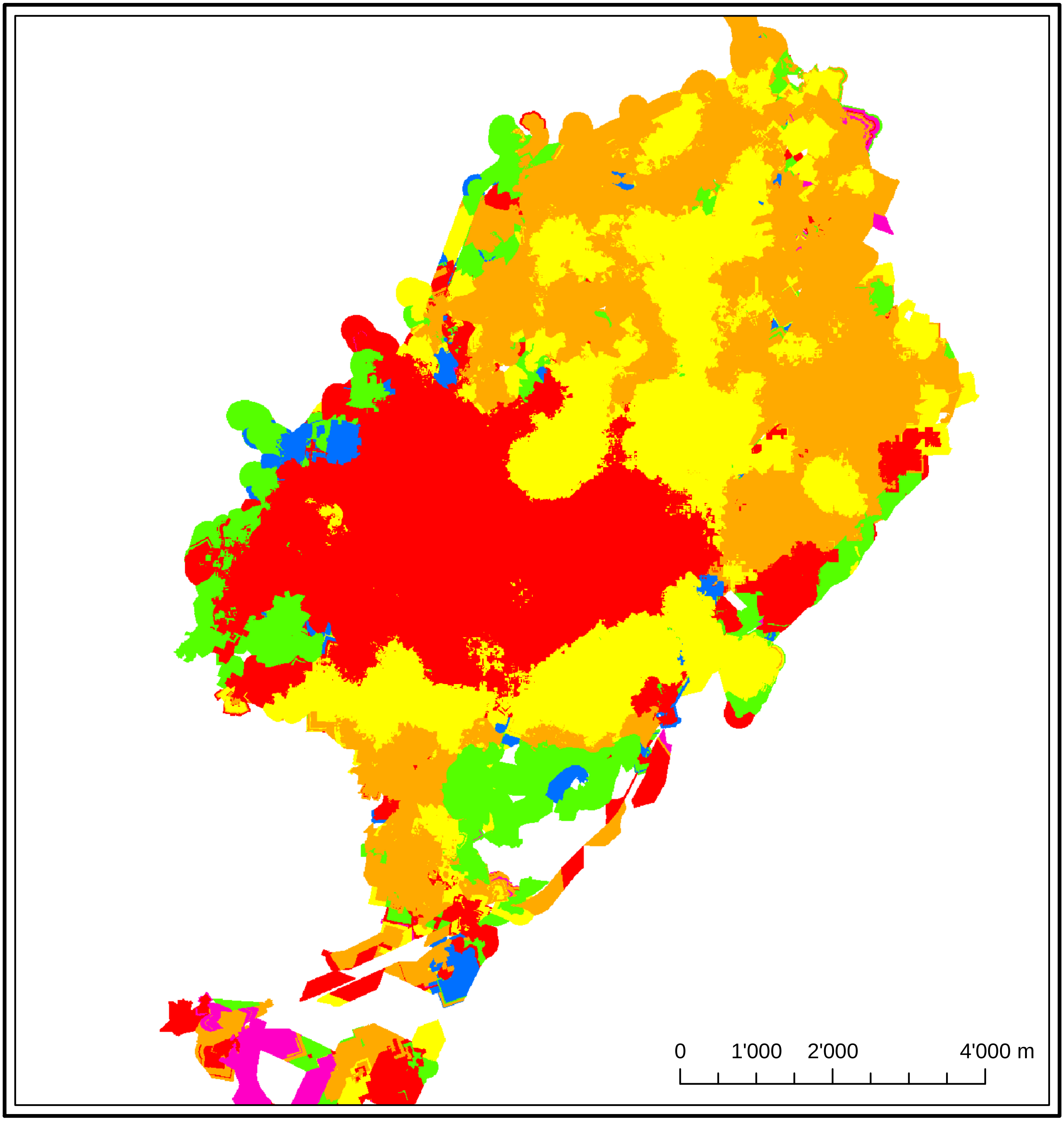
Entropy

	-1.92 - -1.77
	-1.76 - -1.72
	-1.71 - -1.66
	-1.65 - -1.59
	-1.58 - -1.48
	-1.47 - -1.16
	-1.15 - -0.128
	-0.127 - 0

Network entropy, h = 400m

Comment:

T.Produit Master Thesis SSIE - LaSIG - EPFL



Legend

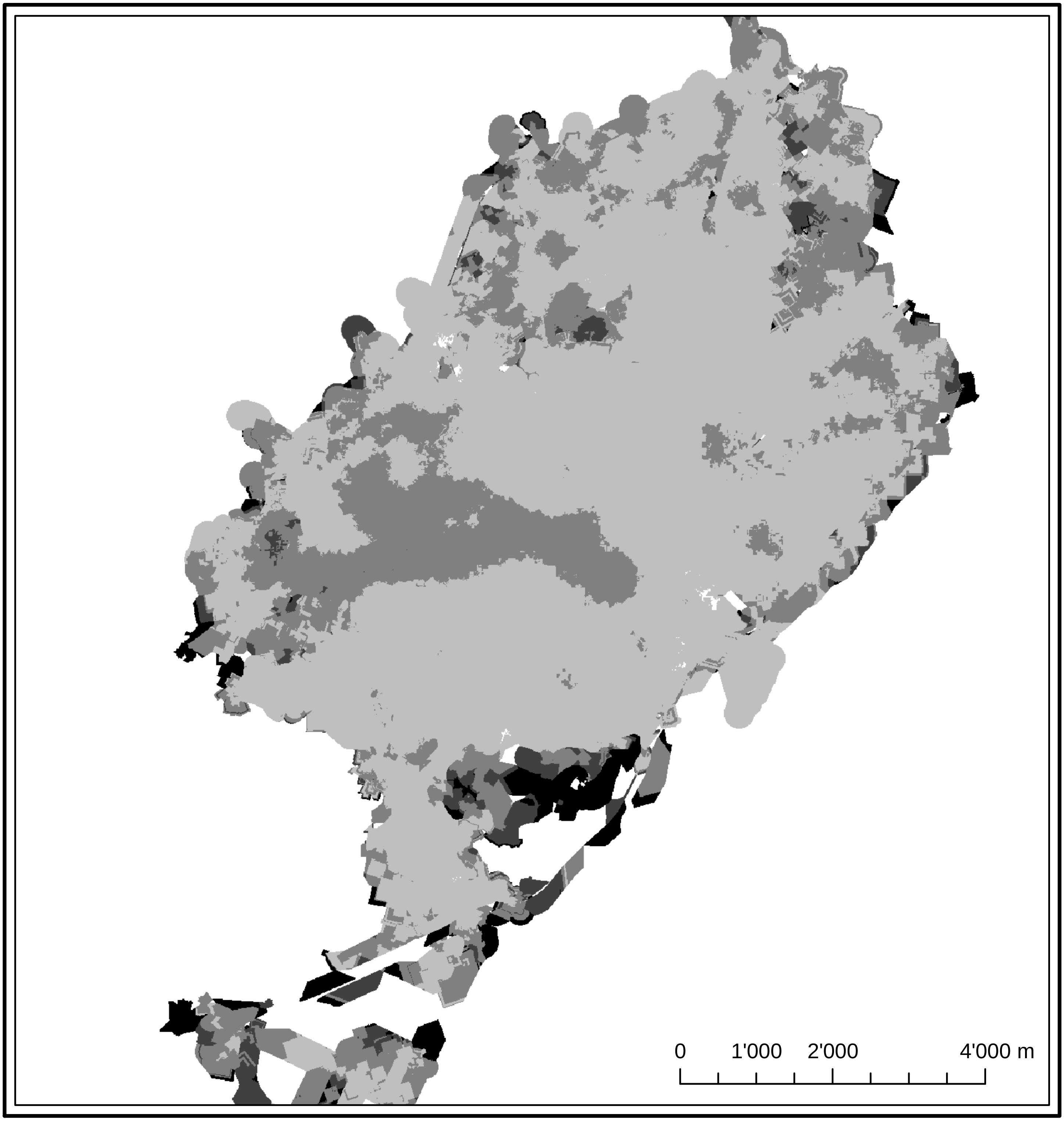
- Others
- Detail stores
- Restaurant and reception
- Prof. or IT activities, services
- Wholesale business
- Public asministration
- Community activities
- No activities

Network diversity, h = 400

Comment:

Predominant specie

T.Produit Master Thesis SSIE - LaSIG - EPFL



Legend

Percentage of the predominant

- 0% - 20%
- 21% - 40%
- 41% - 60%
- 61% - 80%
- 81% - 100%

Network diversity, h = 400

Comment:

Ratio of the predominant specie

T.Produit Master Thesis SSIE - LaSIG - EPFL

5.2.8 Other centralities

For the other centrality indexes the results are less convincing. They mostly represent the density of the edges. The next page illustrates it. It represents the network and a local closeness. Then, the KDE and the NDE are applied, first, simply to the network, and next, to the network with a value of centrality.
One can see that both the NDE and the KDE are not changed by the centrality. The same thing appears with the other centralities. Indeed, the distribution of the betweenness centrality is quite different than the other (Fig 5.20 and Fig 5.16, page: 73). It distribution according to the length of the edges is not the same as well. The Fig 5.21, page: 86 shows these distributions, one can see that the global betweenness is logaritmic. Moreover, the long edge tend to have high values and inversely. This tendency does not appear in the other centrality indexes. All this material explains that the results of density (both KDE and NDE) for the *global betweenness* are better than the others.

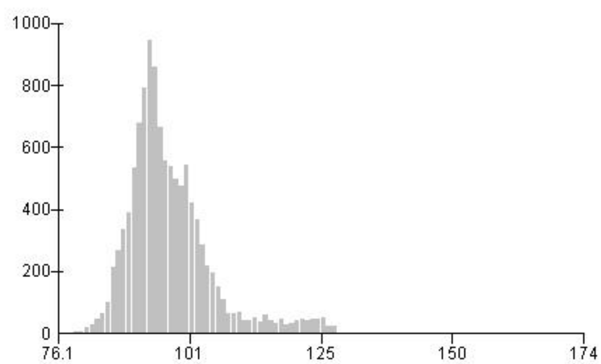
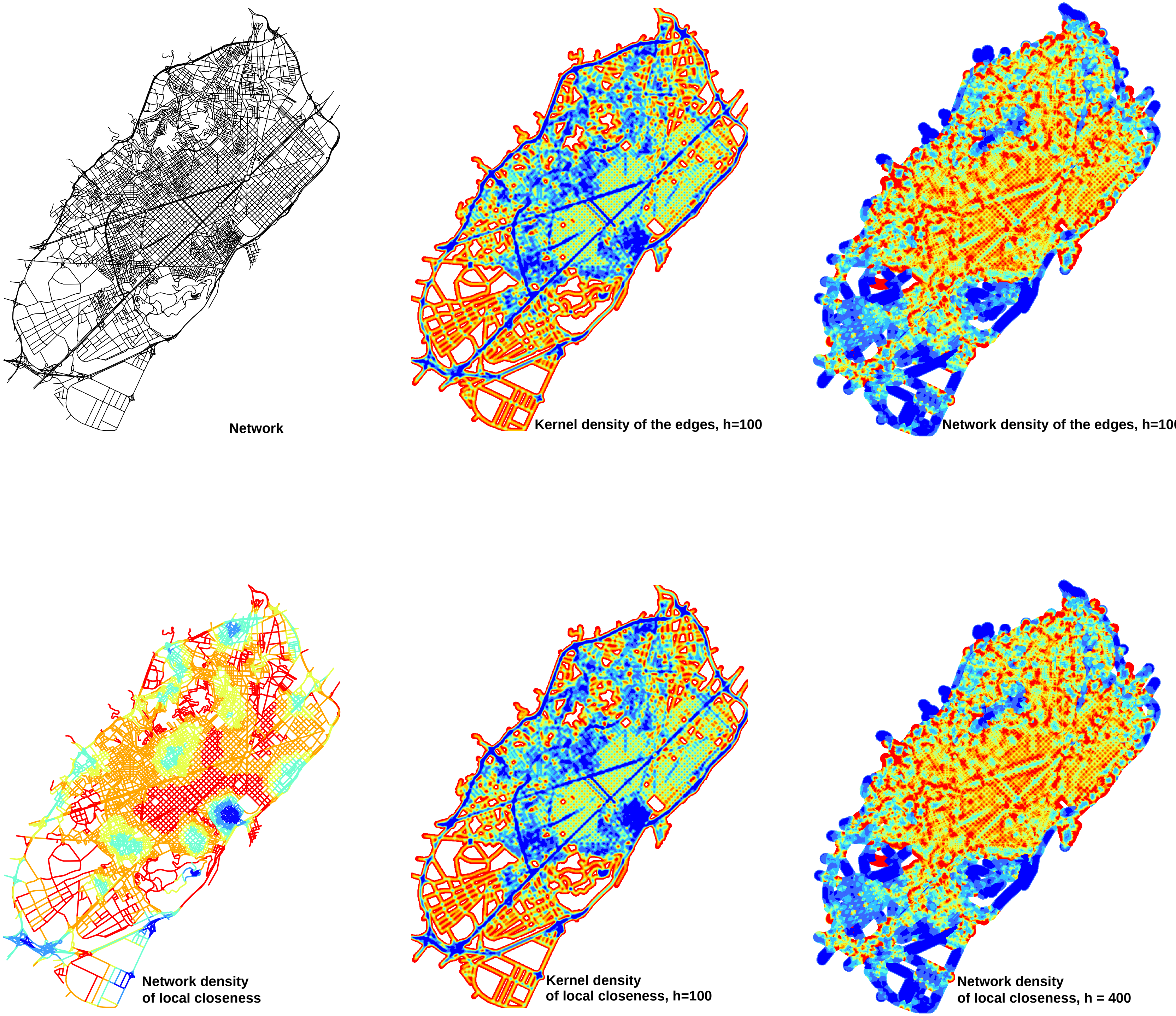
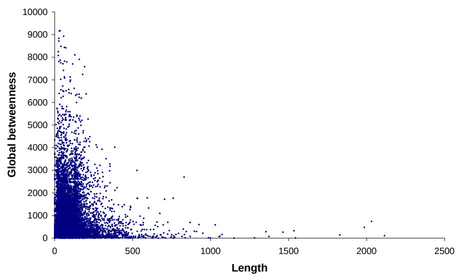
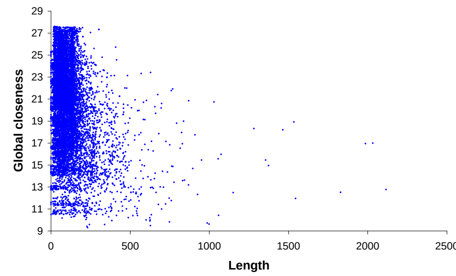


Figure 5.20: Distribution of the edges values for the local closeness centrality, h=1600

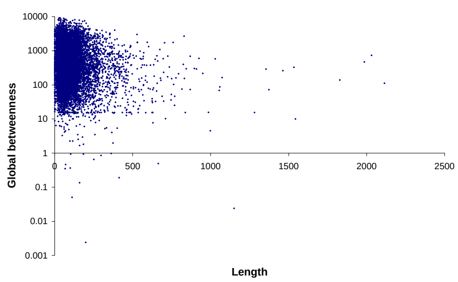

Network
Kernel density of the edges, h=100
Network density of the edges, h=100
Network density
of local closeness
Kernel density
of local closeness, h=100
Network density
of local closeness, h = 400



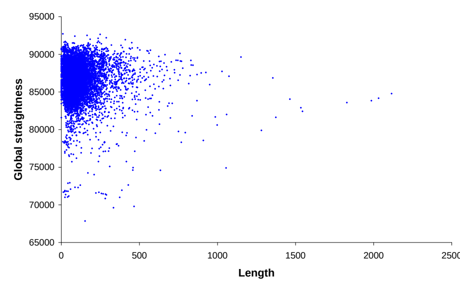
(a) Global betweenness



(b) Global closeness



(c) Global betweenness with logaritmic scale



(d) Global straightness

Figure 5.21: Distribution of the centrality index according to the lengths of the edges. For the *global betweenness* a logaritmic scale has to be used to reach the same kind of distribution.

5.3 Results

5.3.1 Expected correlation

What does mean centrality in a city? One can says that a location is *central* if the urban routes or urban flux tend to cross it. Such a location should be very attractive for managers which want to increase the visibility of their company. A way to check if the centrality indexes match with land-use is a correlation analysis. Such an analysis will be done in a second step of this project by the HSL. Nevertheless, here a first visual analysis is done to anticipate these results.

NDE of betweenness globality and NDE of activities, h=400

Visually, it is not possible to give a trend between these two indicators. Nevertheless, a positive correlation can be expected. Indeed, both the NDE of *activities* and NDE of *global betweenness* become from the same shortest path tree. Moreover, values near or on the network are higher than the others in both case due to projection of the grid points.

NDE of betweenness globality and diversities, h=400

For such a visual analysis, both the euclidean and the network approach can be used without change. Thus, it seems that the high densities of centrality generaly match with low values of diversity and inversely. Always assuming that betweenness centrality match with big streets, these kind of streets seems to be less attractive for some species of activities.

Other NDE of centrality

The other NDE of centrality, above all, exhibit the density of the edges, highlighting the dense area of Barcelona. Such dense area match often with high densities of activities. Nevertheless, whithout a more rigorous analysis there is no way to extract really a trend.

Barcelona's geometry

One can see that Barcelona's network is particular. The *Eixample garden city* with is repetitive pattern, has a important impact on the results. A change in these patterns, like a diagonal, immediately change the centralities and network densities. Such patterns may increase the correlation as well.

The visual analysis of these maps is tricky. First, its author tends to find what he searches, then the scale of the patterns dont exhibit big patch of values easy to evaluate. So, only a genuine correlation analysis will give reliable trends.
This chapter presented the indexes calculated with the network approach, in the next one these indicators will be evaluated and criticized.

Chapter 6

Discuss and prospect

In the previous chapter, the indexes calculated with the network approach were presented, and their results for the city of Barcelona discussed. Here, the indexes are evaluated and critiziced. On this basis prospects are introduced.

6.1 Results and Review

Along this report, new indexes have been presented: the *network density* of activities, the *network density* of edge with centrality indexes as values of "population", and finally euclidean and network diversity.

6.1.1 Properties of the shortest path tree

The SPTs are the basis for the indexes developped in this report. Here are some of is characteristics.

1. Define an area shorter than the euclidean one.
2. Follow the favourite direction.
3. Sensitive to the network straightness.

6.1.2 Indexes

Total length of the SPTs

This value is a good indicator of the density of the network, unlike the NDE, it doesn't take into account the network distance of an edge. Its interpretation is easy

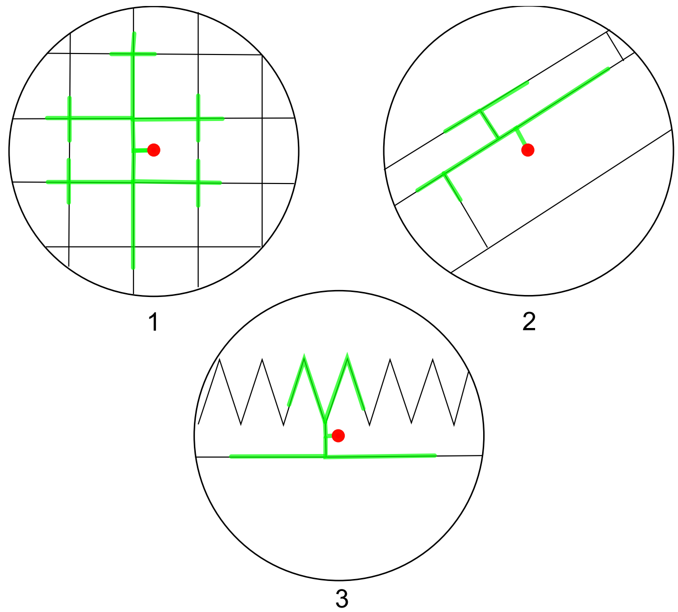


Figure 6.1: Some properties of the shortest path tree

to do and it has a unit [m]. It gives smooth results but have no strong relationship with the centralities. Mainly, it helps to understand the other parameters.

Number of activities

This indicator allows to highlight the economical center of a city. It is not directly linked to the centralities as well, but can help to understand how the economical city is built.
From these two last values, one can create a non-distance weighted density of activities. It only needs to divide the number of activities by the total length of SPTs to obtain a value with a physical meaning and a unit of [activities/m].

Network densities

In order to evaluate the *network density* approach, it is compared with the standard *kernel density estimation* which is a well defined index with a mathematical background.
In the next few lines, the *network density of activities* and the *network density of edges* are discussed. However, the general NDE approach has some weakness. Thus, the strongest limit of the *network density estimation*, is the non-uniform space cre-

ated by the *shortest path tree*. Indeed, the *kernel density estimation* is founded in a uniform euclidean space. In this space, the values of density have a physical meaning. Nevertheless, the NDE, is a KDE applied to a non-uniform space, thus the calculated values have to be used like indicators, because they loose their physical meaning.
Moreover, once a population value is used both KDE and NDE becomes even more difficult to interpret.

Network density of activities

The NDE of activities gives good results even if it is strongly affected by the edge effect. A spatial distribution scale factor appears between this index and the KDE of activities, due to the difference between the catchment area of a circle with a 400m radius and the 400m shortest path tree for example. This scale factor principaly changes with the straightness of the paths. Generally, the spatial distribution of the values of the NDE of activities with a 400m bandwith match well with those one of a KDE with a 100m bandwith.
With the edge effects, there is number of irrelevant values. In this case, they are the zero values and the highest values. For a more sensitive analysis, they must be left aside. In this way, the network approach for activities is justified and can gives a new point of view of the distribution of activities in a city.

Network density of centralities

First, the NDE of centralities is based on a simple NDE of edges. One more time, the *edge effects* are important for this index and zero and too high values may be erased. Then, if you focus on a very dense area, as the the inner city of Barcelona. You will see that the area with a lot of short segments don't have the highest density values. Indeed, as the length is used as a value of population, the highest value are shifted onto the locations dense in middle size segments.

The *kernel density estimation* is defined for cloud of points. Nevertheless, its application for edges is not well founded. In *ArcGIS*, the *kernel density* of edges take into account the distance from the grid cell to the edges as input for the K function, but it is not very clear how the lengths appears (if they directly do).
For the *network density estimation*, the distance value is the network distance from the grid cell to the center of gravity of the edge. Then, to take into account the length of the edge, a value of population is used. Thus, in a certain way, the lengths of the

edges appears two times in the calculation: first, directly in the value of population, then, indirectly with the distance from the centroid to the center of gravity.

The network density of global betweeness is the centrality index which gives the better results. It is especially related to the KDE at lower bandwith. However, during the NDE, the global betweenness lose its logaritmic properties.
The other *network densities* of centrality are mainly related to the the *network density* of the edges and thus don't bring additional informations.

The MCA are indexes created in a primal way to characterize the centralities of the network nodes. Then, they are generalized to the segments with a mean of two nodes. The NDE make use of these centralities as a value of population for the *network density* of edges. One can says that the path between the topology of the network and the NDE of centralities is long. Indeed, an indicator of indicators is calculated. During these steps, there is a risk that informations are loosed. That's what strongly happens for most of the centrality indexes except the global betweenness.

Diversity and dominancy

Shannon's entropy allows to take into account the number of species and their ratio. Roughly, the differences between the network and the euclidean approach appear not important. The distinctive patterns stays in the same place. Nevertheless, the network approach avoid the "disc effect" and so, give a new point of view of the diversity. This index allows to discern patch of high and low diversity which are related with particular area of the city as commercial or residential zone.

To facilitate the interpretation, this network entropy is completed with a value of dominancy, namely, the dominant specie and its ratio. However, in order to avoid a superficial analysis, a good knowledges of the city is required. Secondly, the diversity is an indicator easy to understand and easy to apply. Nevertheless, its background is demanding. Indeed, the sorting of the specimens into species has a great impact to the results. This classification have to be done carefully and according to the research field.

6.1.3 Other aspects

Comparison of the values

One can see, that the KDE and NDE values are not in the same scale of values. First, a problem is the *kernel density estimation* black box function of *ArcGIS*. Indeed, it's difficult to know what it really does and how it treat the values. The only things that are known, are the inputs and outputs.
Moreover, the NDE is physically not well defined. Thus, the comparison should be done on the spatial and/or statistical distribution of the value and not directly on the calculated values.

Study scale

The study scale can impact the result, it is possible that, at the scale of the city, two factors have no correlation. However, these two factors at the scale of the inner city could have a higher correlation. This is the work of the user, following his goals, to choose and apply the better scale. For example, in Barcelona, the global betweenness highlights the big streets, such streets may be not very attractive because the traffic. At the the scale of the inner city, global betweenness should be the attribute of more attractive streets.

Correlation analysis

A correlation analysis between the NDE of the centralities and land-use indicators as the network diversity or the NDE of activities will be done in a second step of this project. It is necessary to warn that these results can be biased by the *shortest path trees*. Indeed, NDE indicators are created with a same bandwith and thus, based on the same *shortest path tree* which may rise unexpected correlation.
Moreover, with the network approach raster density values are always higher near the edges. To avoid such biased results, it may be more rigorous to mix *network* and *euclidean* based indicators in the correlation analysis.

K function

The will for this report was to be as close as possible of a KDE, using a network approach. It leads to the use of the same kernel function for the weighting of the distance. Some test were done to ensure that the results are not much sensitive to this function (a parabolic function was used for the comparison). Actually, this report agree with the conclusion of other authors that the results are more dependant of the bandwith than the K function.

GIS tool

A GIS tool is provided as appendix for this report. The work at distance prevent regular feedback. Thus, the tool is even in evaluation phase, some iterations between users and programmers may be usefull to develop a tool as close as possible to the HSL needs.

6.2 Prospect

The network approach for events occuring in a city is a novel approach. The previous parts, showed that it asks number of questions. Indeed, the *network density estimation* as presented in this report can be improved in some points.

First, at the scale of the algorithm, it was explained earlier that a grid point is projected onto the only nearest edge of the network. A next challenge, which can smooth the results and be more rigorous, is to take into account the surrounding edges (FIG 6.2 and FIG 6.3, page 95).

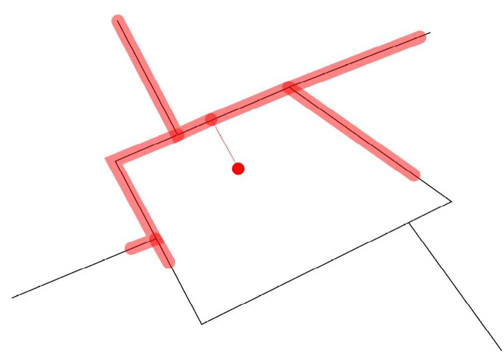


Figure 6.2: First approach, the grid point is projected onto the single nearest edge.

An other limit of the *network densities* approach, is the edge effects which occurs mostly near long and isolated edges. So, a great improvement may be the detection

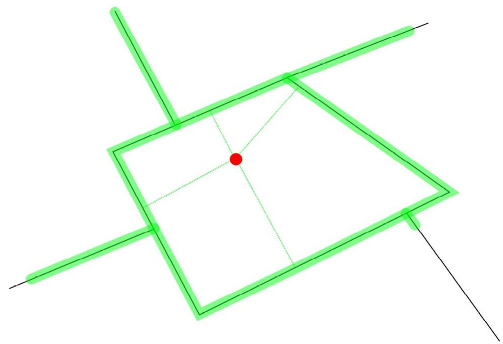


Figure 6.3: Second approach, the grid point is projected onto the surounding edges.

of such geometry to directly erase the inopportune values.

Then, it is difficult to evaluate the results on the basis of a single case study. These indexes have to be applied with different bandwiths to different cities network. Indeed, more studies will allow to have a deeper understanding of the results. In the same way, it would be interesting to apply this network approach to other events than activities.
An expectation for this tool, its that it will increase the correlation between the centrality indexes and indicators of land-use. Along this report, only a visual analysis is done. A strict correlation analysis will more accurately measures how much centralities are linked to other indexes.

Finally, Okabe develop three kernel functions for the estimation of density on a network. These functions are mathematically defined to provide unbiased results [24]. This research is probably the one which gives the strongest background to the density estimation on a network. In this way, these functions have to be deeply investigate.

Chapter 7

Conclusion

The densities calculated with GIS make use of the euclidean space. Nevertheless, events occuring in a city tend to follow the urban network. This report has presented a new tool able to create indexes based on distances measured along a network. These indexes are a *network density estimation of activities*, a *network density of edges* with a value of population made from centralities, and a *diversity* of activities along the network.
All these values are calculated on the basis of the *shortest path tree* which represents all the edges reachable from a point for a chosen bandwith.
Basically, the *network density estimation* is a *kernel density estimation* for which the weighting distances are measured along the network rather than in the euclidean space. The network diversity of activities is a *Shannon's entropy index* applied to activities projected onto the network.

Theses indexes are applied to the 11,200 network edges and 166,300 activities of Barcelona. Results shows that the values are related with the economical city. They are able to highlight particular area, as the shopping line, residential zones or the inner city. The less convincing indexes are the network density estimation of the centralities wich are more related to the density of the edges than their values of centrality.
The space delimitated by the *shortest path tree* is sensitive to the favourite directions of the network. In this way, the indexes founded on this space are more related to the urban routes, and thus, urban reallity of city than those based on the euclidean space.

During the application of the *kernel density estimation* to network distances, the indexes loose their physical meaning. Indeed, the *shortest path trees* are non-uniform and not in the euclidean space. Thus, these values are more indicators than real densities, they have to be used in fonction of this limit.
Moreover, the *network density* is even applied to edges with a centrality attribute.

Thus, the values calculated have to be used and interpret carefully.
This tool is still in evaluation, it need a deeper validation by reasearchers used to the centrality index, kernel density estimation, or more generally urban designers.

The *shortest path tree* is created from a raster cell centroid projected onto the single nearest edge of the network. A more complete approach could project this point onto the surounding edges. To complete the evaluation of these indexes and their possiblities, it is even necessary to do a real correlation analysis. Some warning for such an analysis are presibed.

Network in GIS are well estabished and routing applications are actually a common use. Nevertheless, calculation of statistical parameter along the network are even an emerging research field. Indeed, the problems as the linear referencing, computational needs, database of linear events are overcome. The last and first results, making use of *kernel density* applied along a network were presented by Okabe [24] in January this year. He quotes several utilisation of such index in hydrology, urban, biology, utility or communication networks.

Bibliography

[1] L. Anselin, J. Cohen, D. Cook, W. Gorr, and G. Tita. Spatial analyses of crime. *Criminal Justice*, 4:213–262, 2000.

[2] M. Batty. Network geography: Relations, interactions, scaling and spatial processes in gis. *Re-presenting GIS*, pages 149–170, 2005.

[3] M. Batty, E. Besussi, K. Maat, and J.J. Harts. Representing multifunctional cities: Density and diversity in space and time. *Built Environment*, 30(4):324–337, 2004.

[4] A. Bavelas. A mathematical model for group structure. app. anthrop. 7: 16-30, 1948. also, communication patterns in task oriented groups. *J. Acoust. Soc. Amer*, 22:725–730, 1950.

[5] P. Bishop and P. Gripaios. Explaining spatial patterns of industrial diversity: An analysis of sub-regions in great britain. *Urban Studies*, 44(9):1739, 2007.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports-Review Section of Physics Letters*, 424(4-5):175–308, FEB 2006.

[7] G. Borruso. Network density and the delimitation of urban areas. *Transactions in GIS*, 7(2):177–191, 2003.

[8] G. Borruso. Network density estimation: A gis approach for analysing point patterns in a network space. *Transactions in GIS*, 12(3):377–402, 2008.

[9] C. Brunsdon, A. S. Fotheringham, and M. Charlton. Geographically weighted summary statistics – a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26(6):501 – 524, 2002.

[10] Chris Brunsdon. Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. *Computers & Geosciences*, 21(7):877 – 894, 1995.

[11] R. Cao, A. Cuevas, and W. Gonzalez-Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, 17(2):153–176, 1994.

[12] Alessio Cardillo, Salvatore Scellato, Vito Latora, and Sergio Porta. Structural properties of planar graphs of urban street patterns. *Physical Review E*, 73(6, Part 2), JUN 2006.

[13] Robert Cervero and Kara Kockelman. Travel demand and the 3ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199 – 219, 1997.

[14] P Crucitti, V Latora, and S Porta. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3, Part 2), MAR 2006.

[15] EW Dijkstra. A note on two problems in connexion with graphs. *numerische Mathematik*, 1(1):269–271, 1959.

[16] VA Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14:153, 1969.

[17] L.D. Frank and G. Pivo. Impacts of mixed use and density on utilization of three modes of travel: Single-occupant vehicle, transit, and walking. *Transportation Research Record*, pages 44–44, 1994.

[18] LC Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[19] AC Gatrell, TC Bailey, PJ Diggle, and BS Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21(1):256–274, 1996.

[20] B. Hillier and J. Hanson. The social logic of space. 1984.

[21] B. Jiang and C. Claramunt. A structural approach to the model generalization of an urban street network. *GeoInformatica*, 8(2):157–171, 2004.

[22] H.A. Karimi. *Handbook of Research on Geoinformatics*. Information Science Reference, 2009.

[23] H.J. Miller. Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures. *Geographical Analysis*, 31(2):187–212, 1999.

[24] A. Okabe, T. Satoh, and K. Sugihara. A kernel density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographical Information Science*, 23(1):7–32, 2009.

[25] A. Okabe and I. Yamada. The k-function method on a network and its computational implementation. *Geographical Analysis*, 33(3):271–290, 2001.

[26] G.I.P. Ottaviano, D. Pinelli, C.J. Maignan, and F. Rullani. Bio-ecological diversity vs. socio-economic diversity: A comparison of existing measures. 2003.

[27] S. Porta, P. Crucitti, and V. Latora. The network analysis of urban streets: A primal approach. *Arxiv preprint physics/0506009*, 2005.

[28] S Porta and V Latora. Correlating densities of centrality and activities in cities: the cases of bologna (it) and barcelona (es). 2007.

[29] S. Porta, V. Latora, F. Wang, E. Strano, A. Cardillo, S. Scellato, V. Iacoviello, and R. Messora. Street centrality and densities of retail and services in bologna, italy. 2007.

[30] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853 – 866, 2006.

[31] J. Scheurer, C. Curtis, S. Porta, and E. Foundations. "spatial network analysis of public transport systems: developing a strategic planning tool to assess the congruence of movement and urban structure in australian cities". In *Australasian Transport Research Forum, Melbourne*, pages 25–27, 2007.

[32] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[33] BW Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[34] EH Simpson. Measurement of diversity. *Nature*, 163(4148):688, 1949.

[35] S.H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[36] S. Taneja. Technology moves in. *Chain Store Age*, 75(5):136–138, 1999.

[37] M. Thurstain-Goodwin and D. Unwin. Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. *Transactions in GIS*, 4(4):305–317, 2000.

[38] WR Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, pages 234–240, 1970.

[39] J.R. van Eck and E. Koomen. Characterising urban concentration and land-use diversity in simulations of future land use. *Special Issue Paper*, 2007.

[40] I. Yamada and J.C. Thill. Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158, 2004.

Chapter A

User Manual

With this part you will learn how to install the different programs which will allow to compute the network density estimations and diversities. Then the second section will explain how to use the provided tool.

A.1 Installation

As explained earlier, the algorithm for the network calculation, uses several programs. Here, you will find the steps to install all this programs.
The tool has been created at the beginning for ArcGIS 9.1. Nevertheless, most of the spatial operations are done in PostGIS which is a spatial extension for PostgreSQL. The central algorithm is written in Python 2.1 which is one of the languages tolerate by ArcGIS.

A.1.1 ArcGIS 9.1

We make the assumption that ArcGIS 9.1 is already installed. Thus, Python 2.1 was installed in the same time.

A.1.2 Python 2.1

This step is not necessary if ArcGIS 9.1 is already on your computer. Double click on "Python-2.1.3.exe"

- The path for Python is usally C:/Python21
- **Choose**: Yes, make backups

- **Select Components**: check all boxes
- Start the installation

In Start/All Programs/Python 2.1 should appears IDLE(Python GUI) which allows to edit and run python scripts.
A new UNWISE.EXE is given: just drop it over the old uninstaller, which should be in C:/Python21/UNWISE.EXE unless you chose a different directory at install time.

A.1.3 PostgreSQL 8.3

This database can be downloaded for free in its web site[1]. The package contains a Stack Builder which gives access to the package of PostGIS.
Double click on the "postgresql-8.3.7-1-windows.exe":

- **Installation Directory**: let the proposed file in the program files
- **Data Directory**: let the proposed file in the program files
- **Password**: Usually "postgres" is used
- **Port**: 5432
- **Locale**: [Default locale]

The installation begins, it should take no more than 5 minutes. Choose "Launch Stack Builder at exit" and then "Finish"

The Stack Builder give access to some extensions for PostgreSQL:

- **Choose the installation.** Normally: PostgreSQL 8.3 on port 5432,
- In **spatial extension** choose the PostGIS version, for this work version 1.3.5 was used,
- Choose one of the links proposed. Then the package will be downloaded in a "Temp" file (you can choose an other folder if you want to keep a backup of the package).

[1] http://www.postgresql.org/

A.1.4 PostGIS

Begin the installation of PostGIS (from the Stack builder it is automatic, else double Click on "postgis(...).exe"

- Both PostGIS (red) and Create spatial database (blue) have to be checked
- **Destination folder**: should be the path where you install PostgreSQL (./Program Files/PostgreSQL/8.3)
- **Database Connection Information**:
  - **User Name**: usually "postgres"
  - **Password**: is the one you choose before: usually "postgres"
  - **Port**: the same as before: 5432
- **Spatial Database Information**: Database Name: postgis

Probably an error will occur (Error opening file for writing...). Choose "Ignore".

Try to use your database. Start/All Programs/PostgreSQL 8.3/pgAdmin III will launch an interface for the database. Your password will be asked when you double click on your server (your password should be "postgres"). Then, open the Databases. Normally, you will find three databases inside (postgres, postgis, template_postgis). postgis is an empty database that you can use for your data if you want. Open postgis database / Schemas / public. Here, in Tables will appear the tables that you will create (two tables are already inside).

A.1.5 Egenix MX Base

Run the .exe given which is specific for Python 2.1.

- Installation Directory: Should be C:/Python21

A.1.6 Psycopg2

Allow to connect Python 2.1 with PostgreSQL.
Just copy libpq.dll and psycopg.pyd and paste them in C:/Python21/DLLs

A.1.7 Install the tools in ArcGIS 9.1

For every tool, two scripts are given. One is done for ArcGIS 9.1, it will allow to have a kind of userfriendly interface. The second one is done for the Python 2.1 shell and editor. This means that they are independant of ArcGIS. You can use them if your version of ArcGIS is not the 9.1 or if you don't have ArcGIS. In this case, the shape files have to be built following the instructions given later with an other GIS.

Add Network Density Estimation Toolbox

- Paste the Network Density Estimation Toolbox.tbx in C: / Program Files / ArcGIS / ArcToolbox / Toolboxes/.
- Paste the folder NDEScripts in C: / Program Files / ArcGIS / ArcToolbox/.
- Launch ArcCatalog.
- Go to Toolboxes / System Toolboxes.
- Somewhere in the list should appear the Network Density Estimation Toolbox that you paste.
- Drag and drop it in the ArcToolbox (ArcToolbox window). The next time you will launch ArcMap this Toolbox should appear and be ready for use.

A.1.8 Last but not least

Here we will add new environment variables.
Go in Control panel / System Properties / Advanced / environment variables. Then, choose Path and edit and paste at the beginning:

```
C:\Program Files\PostgreSQL\8.3\bin;
C:\Program Files\PostgreSQL\8.3\lib;
```

A.1.9 Finish

If you haven't had major troubles during these steps everything should be OK to begin working.

A.2 Work begins

All the scripts that you will run with ArcGIS should open the ArcMap shell (black window). In this window you will read some information about the running, like the path used, the files created... More important, if an error occurs, the text "[enter] to continue" will appear. The line before this message give informations on the kind of error. If you type "[enter]" the calculation will try to continue but probably it will crash, except for minor error like a table column already having this name.
At the end of the process, this same message will appear, the previous lines give informations about the results of the process. Type [enter] and the shell will close.

A.2.1 Ressources

For the map that have been presented in the previous chapter we had:

- 11,222 edges for a total area of 92.65km^2,
- 166,311 activities,
- The projection of the 1,890,000 initial grid points takes 36 min. After this step only 926,539 grid points remain, the others were too far from the network.
- The projection of the 166,311 activities takes 2 min.

The calculation was for a 400m bandwith:

- The computing of the euclidean diversity takes 4h.
- The computing of the network densities and diversities takes 33h.

The computer that has been used is:

Intel(R) Core(TM)2 Quad CPU
Q950 @ 3.00GHz
2.99Ghz, 7.83 GB of RAM[2]

[2]Control Panel/System/General

A.2.2 Create the grid of point

This is done with the script CreateGridPoint.py. This tool is the only one which is specific for ArcGIS. This means that it calls some of the functions of ArcGIS. Double click on CreateGridPoint on your new Toolbox.

- **Path to network**: Choose your network shape with the browser. We use the network to find the max and min coordinates of this shape.
- **Cell size**: This is the cell size or raster size for which you will do all the calculations.
- **Extent**: The extent you choose will be added or subtracted of the max and min coordinates of the network shape. The question is how from the network you want to have raster cells.

An ArcMap shell will open and give some informations on the current steps. At the end type [enter] to go back to ArcMap.

First, the script will create a raster with zero value over your network[3]. You can see this raster in the path that you give. Its name is constraster. Then, this raster is converted in points[4]. The name of the shape file is gridpoint.shp. You can add this shape to your map, if you want.
Note that all the points too far from the network (bandwith size) will be deleted in the next steps.

A.2.3 Check your data

In the next step, we will put all our datas in the database to begin spatial calculation. So, here, we will check that our datas are well designed.

In the network you must have the columns (exactly the same name, but upper or lower case doesn't matter):

edge_id, fnode, x_fnode, y_fnode, tnode, x_tnode, y_tnode, length

Then you have values of centralities.

[3]Spatial Analyst Tools/Raster Creation/Create Constant Raster
[4]Conversion Tools/From Raster/Raster to Point

In the activities, you must have a column species (exactly the same name), which contains an integer representing the species of the activity. Only a specific function species.py[5] is given. This function converts the codi_siscl in the species for the case study of Barcelona. For an other network, you can both modify this function or do a calculation. For example:

- Right click on your activities shape.
- Choose Open Attribute Table.
- In Option choose Add Field, name: species, Type: Short Integer.
- Right click on the heading of your new column and click Calculate Values...

You can do it with Excel in the dbf file as well.

Nothing is important to check in your grid of point, only the geometry will be used.

A.2.4 Create a database

We will put the network, the activities and the grid of point into this database.

- Run: Start/All Programs/PostgreSQL 8.3/pgAdmin III
- Double click on PostgreSQL 8.3 (localhost:5432)
- Right click on Databases and choose new database
- **Name**: Choose a name for your database
- **Owner**: usually it is postgres
- **Encoding**: WIN1252
- **Template**: template_postgis, means that you will copy the design of template_postgis
- Do not change the other parameters and click OK

Your new database should appear in Databases. If this is not the case, you can refresh the object with the button with a red and green arrow.

[5] /MasterTProduit / Scripts / withoutAG / speciesDBwa.py

A.2.5 Import your shape in the database

You should have three shapes: the network, the grid of point, the activities. For convenience, put them in a same folder near the root like C:/temp/data/. The function insertDB is done to select the shapes, convert them in SQL request and put them into the database via the shell.

With ArcGIS 9.1

Double click on insertDB. Choose a shape file (network.shp, gridpoint.shp or activities.shp) with the browser. Give the name of your database, your user name (postgres), the host (localhost) and the password (postgres).
The shell (black window) will open. If everything is ok, you will have to click [enter] only one time on this shell at the end of the process. Otherwise an error message should give you the reason of the crash. Without trouble you will find in the same folder as your data the new files .sql and a new table in your database.

Without ArcGIS 9.1

Edit the script insertDBwa.py with Python GUI, at the beginning of the file change the parameters in green to fit with what you want.

With the shell

Use this solution, if the two others crash or if you want to use your spatial database for other goals.

- Open the shell (Start/Run and write cmd and OK).
- Go to the root of the binary files of PostgreSQL

  ```
  cd C:\Progra~1\Postgr~1\8.3\bin
  ```

- Run this kind of command:

  ```
  shp2pgsql -I C:/data/myshape myshape > C:/data/myshape.sql
  ```

  This line asks to use the function shp2pgsql.exe for the shape file network in the folder data and convert it in a sql request (you can edit it with your favorite editor) in the path given.

- And then this one:

  ```
  psql -d dbname -f C:/data/myshape.sql -U postgres
  ```

  With dbname: the name of the database and postgres your user name. This line runs the sql request in the database using psql.exe.

Check your data in the database

You must have tables network, activities and gridpoint (exactly these names, else change them). If you want to see what your data looks like: Open pgAdmin III, go to your database, refresh and check your table. Using the tool in the top of your window Run your own request (Yellow pencil), you have access to an SQL interface. Here you can write something like:

```
SELECT * FROM gridpoint LIMIT 100;
```

This line will show you the 100 first lines of the table gridpoint. For more complex requests, look at the documentation of PostgreSQL.

A.2.6 Clean the network

With ArcGIS 9.1

This function will clean the table network. Because at this time some edges have the same fnode and tnode with a zero length. (Don't know why!). These values will induce a crash.
Double click on clickNet and only give parameters of connection.

Without ArcGIS 9.1

Edit cleanNetwa.py and change the parameters.

A.2.7 Project activities on the network

This function (projAct) find the nearest edge for an activity and compute the distance and the geometry of this point. It give as well the linear measure of the point on the edge. (The projection is at 50m of the beginning of the edge).

With ArcGIS 9.1

Double click on projActivities in the Toolbox.
In Distance Max. give the maximal distance for an activity to be projected on the network (Fig A.1). It should not exceed the bandwith that we will use for the NDE. Nevertheless, choose this parameter carefully, because this value will be the search radius for the edges around the activity. Then the distance to these edges are calculated to keep only the edge with the minimal distance. So, if this value is big, the running time will increase.
Give the name of your database, your user name (postgres), the host (localhost) and the password (postgres).
This function should take not too much time (2min for Barcelona with 200m as radius).

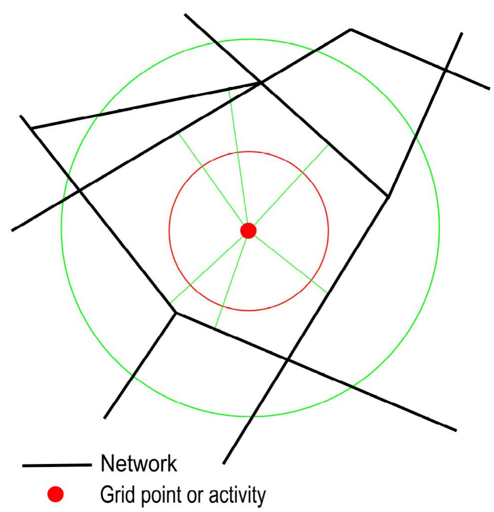


Figure A.1: Distance Max parameter: **Red circle**, represents a distance max. too small, no edges are clipped the point is not projected. **Green circle**, some edges are clipped, the algorithm search the nearest one.

Without ArcGIS 9.1

Edit projActwa.py and change the parameters.

A.2.8 Project the grid of point on the network

This function (projGrid) finds the nearest edge for a point of grid and compute the distance and the geometry of this point. It also give the linear measure of the point on the edge. (The point is at 50m of the beginning of the edge). More columns are added than for the function projAct, like the grid point and its projection coordinates.

With ArcGIS 9.1

Double click on projGrid in the Toolbox.
In Distance Max. give the maximal distance for a grid point to be projected on the network. It should not exceed the bandwith that we will use for the NDE. Nevertheless, choose this parameter carefully. Because this value will be the search radius for the edges around the activity. Then, the distance to these edges are calculated to keep only the edge with the minimal distance. So, if this value is too big the running time will increase.
Give the name of your database, your user name (postgres), the host (localhost) and the password (postgres).
This function should take a bit more time (40min for Barcelona with 200m as radius).

Remark: Here, it could be interesting to see what the projected activities, projected grid point, activities or grid point looks like after this process. Indeed, the activities or grid points too far of the network have been deleted. So, run pgsql2shp. This function will convert a table (gridpoint) into a shape file (gridpointDB.shp). First, you have to give the folder in which you want to save your shape. Then, give the name of the table (gridpoint, activities, projpoint, projact) and the parameter of connexion. The shell should open, give some informations about the paths and command lines used. Then, the conversion begin (a lot of X represent the status). At the end, type "[enter]" and check if the shape file appeared in the path.

Without ArcGIS 9.1

Edit projGridwa.py and change the parameters

A.2.9 Create a raster of euclidean diversity

This function will compute Shannon's diversity for the activities inside a choosen radius around the grid point. The script looks at the column species of the table activities to create the diversity (see: 4.5.9).

In ArcGIS 9.1

Double click on Euclidean Diversity in the Toolbox. For the radius, choose the radius in which you want to compute the Shannon's diversity. Then fill the connection parameters for the database.
The ArcMap shell should open and give the status of the work (percentage).

Then, run pgsql2shp. This function will convert a table (gridpoint) into a shape file (gridpointDB.shp). First, you have to give the folder in which you want to save your shape. Then, give the name of the table and the parameter of connexion.

Now, you have to convert these points into a raster file. In ArcMap choose the Tool: Conversion Tools/To Raster/Feature to Raster. Choose the Input features (gridpointDB.shp), the Field (diveucl), change the name of your Output raster (ex: divEucl200), give the Output cell size wich should be the same that the size you choose to create the grid of point (10m).
The raster of euclidean diversity appears in ArcMap.

Without ArcGIS 9.1

Edit diverstiyEuclideanDBwa.py, change the parameters and run the script. You will obtain the table gridpoint with a new column diveucl. You can now convert the table into a shape file, by editing and running pgsql2shp.py. Convert the shape in a raster with an other GIS.

With the shell

Use this solution, if the two other crashs or if you want to use your spatial database for other goals.

- Open the shell (Start/Run and write cmd and OK).
- Go to the root of the binary files of PostgreSQL

```
cd C:\Progra~1\Postgr~1\8.3\bin
```

- Run this kind of command:

  ```
  pgsql2shp -f C:/data/gridpoint -P postgres -u postgres dbname tablename
  ```

  This line asks to use the function spgsql2shp.exe to create the shape file gridpoint in the folder data from the table tablename in the database dbname with the password (-P) postgres and the username (-u) postgres

A.2.10 Create the raster of network density of centralities and activities

This script has three goals:

- Compute diversity of activities on the network
- Compute Network Density of activities on the network
- Compute Network Density of centralities

This calculation is time and memory consuming. So, before running it, think to restart the computer and open only the necessary softwares. As ArcGIS is memory consuming, it is better to use the method "Without ArcGIS 9.1"

With ArcGIS 9.1

Double click on NDE_DivNet in the *toolbox*.
The first parameter that you have to fill is the bandwith. It is the length of the SPT and the bandwith for the densities as well.
Then the first "check box" is for the computation of the network density of centralities. If you check it you have to fill the next input with the name of the columns of centralities like this:

```
bet_glo;clo_glo;str_glo
```

Remark: if you want to compute a simple network density, you have to create a new column in the network with this SQL request:

```
ALTER TABLE network ADD column one INTEGER;
UPDATE network SET one = 1;
```

Then, put this column "one" in the name of the centralities.
After that, the two last "check box" give you the choice of computing the network density of activities and diversity of activities.

The first thing that the script will do is check the column's name, you will read an error message in the shell if one of them does not exist.

This calculation can take a long time, if the computer isn't near the burning point, a message in the shell should give you the progress of the work. Now, you deserve to have a break and a beer, come back in some hours/days to type "[enter]" in the shell at the end.
To create the raster, look at the steps explained in section A.2.9. For the name of the column, please look at section D.3.

Without ArcGIS 9.1

Edit NDE_DivNetwa.py and change the parameters.

Chapter B

MCA user Manual

The Multiple Centrality Assessement is the approach that the HSL has developed to give centrality values to the edges of a network. In this way an *ArcGIS* tool is available. This part will explain how to create such index for a network.

B.1 Installation

First, this user manual is done for *ArcGIS 9.1* users, for other versions the tool have to be tried.

1. Double click in the McaExtensionSetup.msi to begin the installation. A new toolbar will appear (Fig B.1).

Figure B.1: The nine new tools of the MCA extension

B.2 Creation of a centrality weighted network

1. The next steps will clean the network to create a topology allowing the MCA algorithm. Click on the second tool, a window appear. Select your network and choose a name and a path to your new shape file. Depending the resolution of your shape choose a buffer radius to search the disconnections (two line close one each other will be merged).

2. Click on the third tool a window appear. Select the previous network that you have created and choose a name and a path to your new shape file. Depending the resolution of you shape choose a buffer radius to search the disconnections (two line close one each other will be merged).

3. Click on the fourth tool a window appear. Do the same steps as before.

4. Click on the fifth tool a window appear. Do the same steps as before.

5. Click on the sixth tool a window appear. Do the same steps as before. Now the network, is clean enough to apply, the MCA.

6. Click on the first tool, a window appear. Select the previous network that you have created and choose a name and a path to your new shape file which will be the **nodes** of the network.

7. Click on the seventh tool, choose the last network that you have created and then select the shape of your nodes. The tool create **the connectivity shape**.

8. Right click on this last shape, choose Open Attribute Table. Then, in Option choose Export, choose the name and the path to this table.

9. Edit this dbf file with Excel, then save it as a txt file.

10. Edit this txt file with your favourite editor and delete the first line. The delimiters have to be space. In the number, the decimal delimiters are dots ([.]).

11. Put CLI.exe: the C++ software which create the centrality index, in the same folder than your data.

12. Double click on CLI.exe. First, give the name of the txt file as network.txt, then choose number of locale scale that you want for the closeness centrality, type [enter]. This program should create txt files with centrality values for the nodes and the edges. The values of the edges are the means of the value of two nodes.

13. Create a new Excel file, paste the columns of the edges centrality in this file, add a line with titles for the columns. Save it as a dbf file.

14. Come back to ArcGIS. Add the dbf file. Right click on the shape of connectivity, and choose Joins and Relates and then Join...

15. There, you have to choose: the joined table which contains values of centrality and the ID's on which the join will be done (EDGE_ID)

16. Right clik on the shape of connectivity, and then in Properties ... In Symbology choose Quantities and then the centrality index that you want to see.

Chapter C

Abreviations

API:	Application Programming Interface
DB:	Data Base, Base de donnée
EPFL:	Ecole Polytechnique Fédérale de Lausanne Lausanne Federal Polytechnical School
GIS:	Geographic Information System
HSL:	HumanSpaceLab, Politecnico di Milano
ID:	Identifiant
IT:	Information Technology
KDE:	Kernel Density Estimation
MCA:	Multiple Centrality Assessment
PoliMi:	Politecnico di Milano
SIG:	Système d'Information Géographique
SPT:	Shortest Path Tree

Chapter D

Material

D.1 Scripts

There are three versions of the scripts:

- In MasterTProduit / ArcGIS9.1 / NDE scripts, you will find the scripts ready to be called by *ArcGIS 9.1*
- In MasterTProduit / Scripts / withAG, you will find the same scripts than the previous ones, but for an editing / modifying use.
- In MasterTProduit / Scripts / withoutAG, you will find the scripts which are independant of ArcGIS.

D.2 Programs

In MasterTProduit / programs, you will find the different package for the necessary programs. Prefer these packages than the ones that you can download from the Web, because these versions are compatible.

D.3 Data

The provided data MasterTProduit / data, is the shape file of the grid of point. Here, are the *translation* of the column names:

- EDGE_ID: the nearest edge;
- TOTALLENG: total length of the SPT;

- NBRACT: number of activities on the SPT;
- DIV_NET: entropy measured on the SPT;
- PRED_SPEC: predominant specie;
- PRED_RATIO: ratio of the predominant specie;
- DENS_ACT: NDE of activities;
- C_(...): NDE of centralities;
- C_ONE: NDE of the edges;
- ACT_TOTALLEN: [number of activities]/[total length] = density of activities;

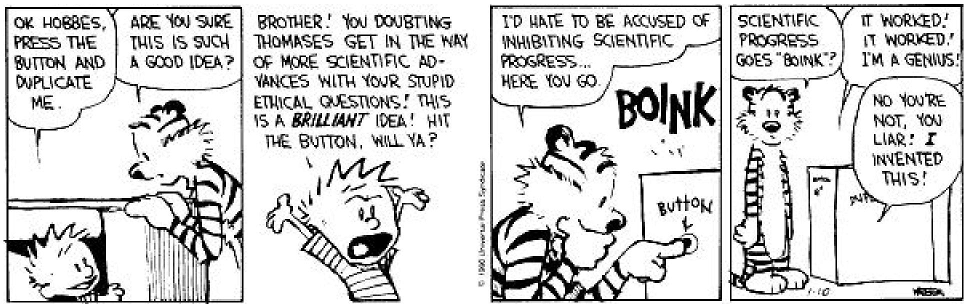
All these values are calculated with a 400m bandwith.

D.4 Report

The report as .tex is given as well, some figures or formula can be usefull (MasterTProduit / report / master).

D.5 Papers

The database of the litterature master.bib and the different paper quoted are in MasterTProduit / report / litterature.



Last figure: Scientific progress goes "BOINK"?
"*Scientific Progress Goes 'Boink'*", page 55, B. WATTERSON

Unformal aknowledgement:

Je voudrais d'abord exprimer ma profonde reconnaissance et mes remerciements sincères et puis aussi exprimer toute ma gratitude à mes collègues de travail de diplôme pour leur support moral... Et c'est pas peu dire que parfois ils ont vraiment dû me supporter: Hugues, Jan, Sylvain, Tristan, Daniel, Laurent et les autres, ainsi que Florence, pour la touche féminine, qui a aussi effectué la relecture.

Toujours dans la catégorie soutient moral et parfois même physique, mes études auraient été ternes comme un dimanche au Poly sans l'unique SATELLITE*, ses cafés, ses BD et ses bières...*

Le support technique n'est pas à délaisser non-plus, ainsi je remercie mon ordinateur (dit "Titine") d'avoir toujours été là pour moi, même durant les moments difficiles où je ne lui ai pas dit que des mots gentils. La même reconnaissance est due à son collègue (dit "l'ordi à Stéphane") car à la main les calculs auraient pris plus de temps et puis il aurait fallu engager toute une école enfantine pour colorier les grilles raster.

Finalement, les produits ESRI, m'ont permis de devenir un expert et un Indiana Jones: expert en installations / désinstallation de ArcView $\rightleftarrows$ *installations / désinstallation de Window (une boucle qui permet de toucher du doigt l'*∞*), et Indiana Jones durant l'aventure qu'était la recherche de la configuration idéale.*