Analysis of the pedestrian network of Lausanne: eccentricity, accessibility and partitionning

Théophile Emmanouilidis and François Bavaud
Department of Geography
Department of Computer Science and Mathematical Methods
University of Lausanne, Switzerland
{temmanouilidis, fbavaud}@unil.ch

Abstract

Indices of sinuosity, eccentricity and accessibility play a major role in spatial planning, quantitative geography, and data-analytic studies. Optimal transportation, path transformation and nodes clustering issues are discussed and illustrated on the complete pedestrian network of Lausanne (27'219 nodes and 31'482 edges).

1 Introduction

The present study exploits and analyses the complete pedestrian network $G = (V, E)$ of the city of Lausanne (135'000 inhabitants, 16 km^2), consisting of a total of $|V| = 27'219$ nodes and $|E| = 31'482$ edges, as managed by a GIS. "Named nodes" $i = 1, \ldots, n$ will in the sequel refer to a subset of $n = 9'613$ nodes consisting of building locations, endowed with an address, inhabited or not. The creation of the network was motivated by the need to determine a fair scheme allowing the 12'016 schoolchildren to be granted free or semi-supported access to the public bus transportation system, depending on the pedestrian distance to the schools, as well as on the age of the children (Figure 1).

This paper investigates the spatial behaviour of the sinuosity, eccentricity and accessibility indices on a real, detailed dataset, consisting of the distances between pairs (shortest-path and Euclidean) and the weights of nodes. Those indices are, beside their descriptive value originating in quantitative geography, most central to planning issues

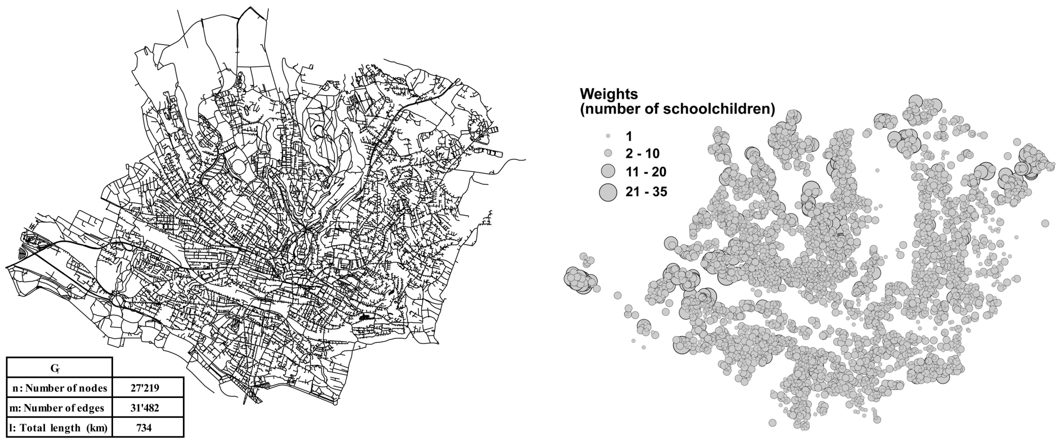

Figure 1: Complete pedestrian network of the city of Lausanne (left) together with the number of schoolchildren, proportional to the nodes weight f (right).

(minimum transportation, minimum cost, short-cuts building) as well as data-analytic issues (geometry of distances, clustering), a fact we shall attempt to exemplify and illustrate in Sections 2, 3 and 4.

2 Distances comparison and sinuosity

Two sorts of distances between pairs (i, j) of nodes can be extracted from the ArcGIS database: the straight or *Euclidean* distances d^e_{ij} and the pedestrian or *shortest path* distances d^{sp}_{ij}. The latter are kept symmetric in this study, thus neglecting the noticeable effect of slope in Lausanne: however interesting, a formally coherent approach involving asymmetric distances seems yet to be constructed. Also, pedestrian walks arguably include the anticipation and incorporation of return trips, thus cancelling the asymmetry.

In addition, nodes are weighted by a distribution f (with $f_i \geq 0$ and $\sum_{i=1}^{n} f_i = 1$), representing in this sudy the proportion of schoolchildren inhabiting node i. Uniform distributions $f_i = 1/n$ yield unweighted graphs.

Analyzing scatterplots of d^{sp} versus d^e yield a pertinent overview of the network connectivity. Linear pattern indicates a good connectivity with $d^{sp} \approx d^e$ (Figure 2), while more diffuse or non-linear patterns indicate a lack of connectivity (Figures 3 and 4). Possible locations for new pedestrian shortcuts can be highlighted by visualizing cases with low d^e and high d^{sp} values (Figures 3 and 4).

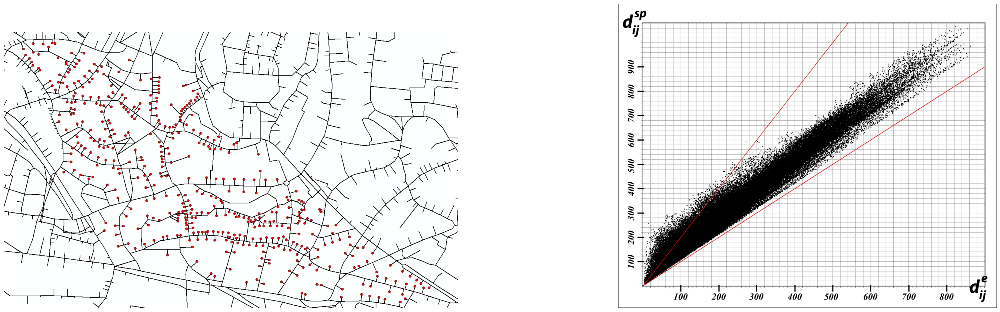

Figure 2: “Rue Centrale” district (441 nodes in the centre of Lausanne, left) and its associated scatterplot (right), displaying linearity. The expected shortest-path distances, as predicted by the Euclidean distances by unweighted linear regression on all pairs of distinct nodes, are $\hat{d}_{ij}^{sp} = 1.17\, d_{ij}^{e} + 46.3$ (in meters), with $r^2 = 0.96$.

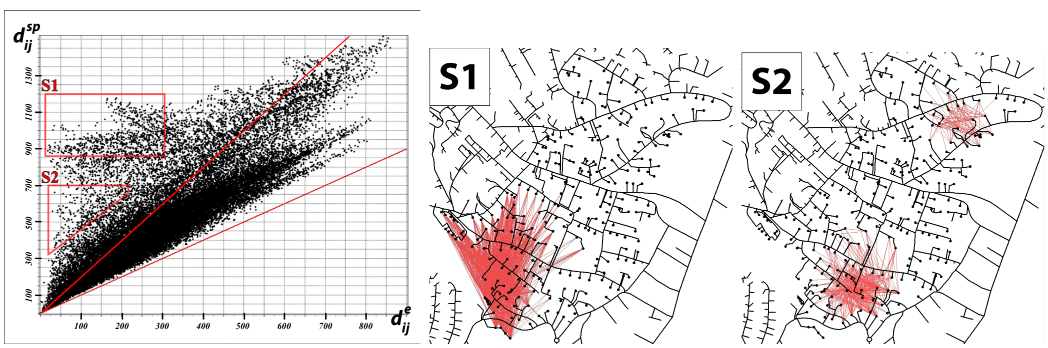

Figure 3: “Plaisance” district (255 nodes, North-East, left). Selecting points in scatterplots can identify possible new shortcut areas. Middle: S1 selection. Right: S2 selection.

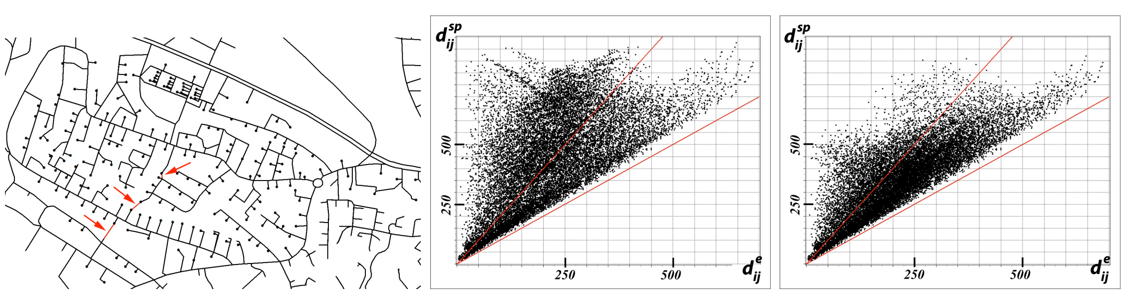

Figure 4: “Montoie ” district (195 nodes, South-West), together with three new pedestrian shortcuts (left). Middle: actual configuration. Right: creating the shortcuts decreases some shortest-path distances, and increases the linearity of the scatterplot.

Figure 5: Subpart of the "Plaisance" district. The average sinuosity Z (high values in red, low values in yellow) detects nodes located in dead-ends or along sinuous streets.

By construction, the *sinusoity index* $\delta_{ij} := (d^{sp}_{ij} - d^e_{ij})/d^e_{ij}$ is non-negative. For instance, the typical sinuosity in the "Rue Centrale" district is 0.17 at large distances, in view of the results of the regression (Figure 2). The *average sinusoity* $Z_i := \sum_{j\in A} \delta_{ij}/n(A)$ (where $n(A)$ is the number of nodes in district A) takes on its largest values at nodes located in dead-ends or along sinuous streets (Figure 5; see also Emmanouilidis 2012).

3 Eccentricities and accessibilities

Let the distances d between nodes together with their weights f to be known. The *eccentricity* e_i and *accessibility* a_i of node i are

$$e_i := \sum_j f_j\, F(d_{ij}) \qquad\qquad a_i := \sum_j f_j\, G(d_{ij}) \tag{1}$$

where d stands for some distance (typically shortest-path d^{sp}, or Euclidean d^e) and $F(d)$ is an *increasing function* of the distance d, defining a travel cost or effort. The average eccentricity $\sum_i f_i e_i = \sum_{ij} f_i f_j F(d_{ij})$ is a measure of the spatial dispersion of the network.

The node i minimizing e_i defines a *centroid*, coinciding with

- the *gravity center* of the district in the squared Euclidean case $F(d^e) = (d^e)^2$
- the *medioid* for $F(d^e) = d^e$ (e.g. Kaufman and Rousseeuw 1990)
- the *p-median* in the shortest-path set-up $F(d^{sp}) = d^{sp}$ (e.g. Hakimi 1965).

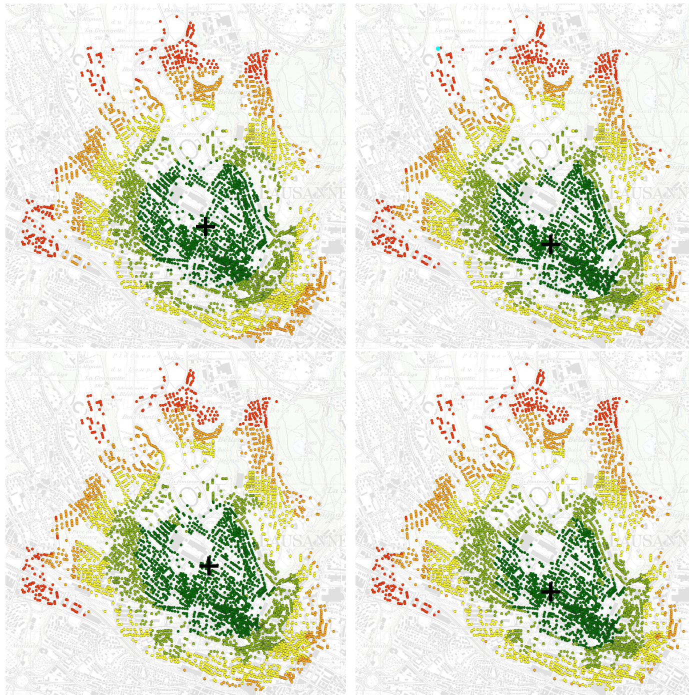

Figure 6: Eccentricities t district of Lausanne (4392 nodes), with penalty functions $F(d)$ defined as d^e (top, left), d^{sp} (top, right), $(d^{sp})^2$ (bottom, left) and $\sqrt{d^{sp}}$ (bottom, right). The cross denotes the corresponding centroid.

Remote locations are effectively taken into account iff $F(d)$ is strongly increasing: when $F(d)$ is flat, little case is specifically made of peripheral contributions. In the context of the optimal locations $L(g)$ of facilities $g = 1, \ldots, m$, such as schools, setting a specific functional form of $F(d)$ hence defines a specific *spatial equity scheme*, paralleling the formal issues involved with *robust estimates of central tendency*. Here the problem consists in minimising the *total cost* facilities

$$C(Z, L) := \sum_{i=1}^{n} \sum_{g=1}^{m} f_i z_{ig} F(d_{iL(g)}) \tag{2}$$

where $Z = (z_{ig})$ is a $n \times m$ attribution or *membership* matrix z_{ig} (obeying $z_{ig} \geq 0$ and $\sum_g z_{ig} = 1$), specifying the proportion of children living at i to be sent at school g located at $L(g)$.

By contrast, $G(d)$ in (1) is a *decreasing* of the distance d, a large value of which characterises easily attainable locations, thus providing an alternative concept of centrality. $G(d)$

is the *distance-deterrence function*, discounting for walks at large distances, and behaving as a *physical potential* in the Gravity model of geographers. Numerical calculations show that long-range potentials (algebraic decay, typically) tend to produce unimodal accessibilities, while short- range potentials (exponential decay, typically) act as a local filter erasing the surroundings beyond some characteristic distance, hence alleviating the difficulties associated with finite observation windows. Moreover, the latter turn out to generate local maxima of accessibility (Figure 7).

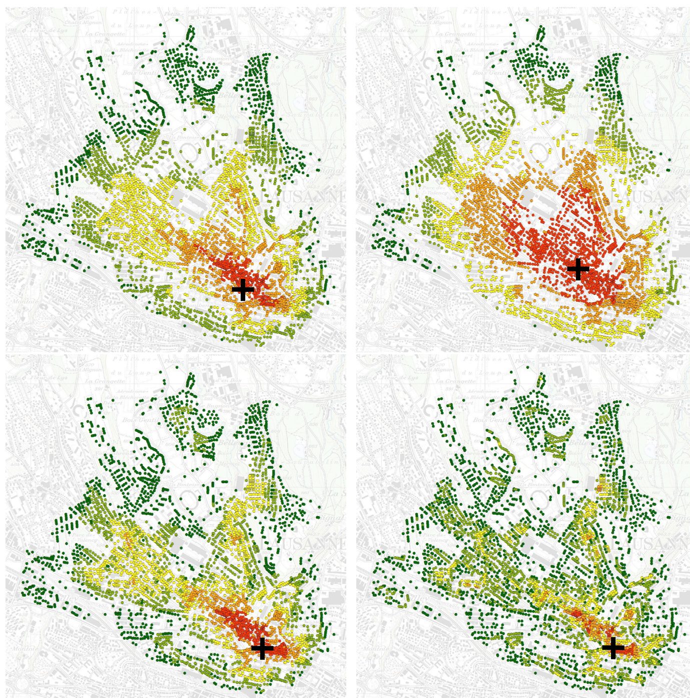

Figure 7: Accessibilities in the North-West district of Lausanne, with distance deterrence functions $G(d)$ defined as $1/d^{sp}$ (top, left, where identical nodes have been discarded), $1/(a + d^{sp})$ where a is the average unweighted shortest-path distance between pairs in the district (top, right), $\exp(-\lambda d^{sp})$ with $\lambda = 0.01$ (bottom, left) and $\lambda = 0.03$ (bottom, right). The cross denotes the node with maximal accessibility. Note the appearance of local maxima of accessibility for the short-ranged exponential potentials (bottom).

3.1 Within-districts and -groups: hard and soft

The eccentricity e_i^A and accessibility a_i^A of node i *in (hard) district* A (where A can vary form the whole city to a single node) are defined as

$$e_i^A = \sum_j f_j^A F(d_{ij}) \qquad a_i^A = \sum_j f_j^A G(d_{ij}) \qquad f_j^A = \frac{f_j\, 1(j \in A)}{\sum_{k\in A} f_k} \ . \qquad (3)$$

Here f_j^A is the straightforward restriction of f_j to some district A of Lausanne (recall the *characteristic function* to be $1(S) = 1$ if S is true, and $1(S) = 0$ otherwise). Figure 6 depicts the eccentricities for the North-West district, and Figure 7 the accessibilities.

Similarly, "soft districts" $g = 1, \ldots, m$, together with associated eccentricities e^g and accessibilities a^g can be defined by replacing f^A in (3) by the "fuzzy profile" f^g defined as

$$f_i^g := \frac{f_i z_{ig}}{\rho_g} \qquad \rho_g := \sum_i f_i z_{ig}$$

obeying $f_i^g \geq 0$ and $\sum_i f_i^g = 1$ as well as $\rho_g > 0$ and $\sum_g \rho_g = 1$, where $Z = (z_{ig})$ is the (generally non-binary) membership matrix in (2).

3.2 Euclidean transformations of the distances

The Euclidean (or not) nature of the network distances d has important consequences: if d is Euclidean, the nodes can be embedded by multidimensional scaling (MDS) in a continuum of dimension at most $n-1$ (e.g. Mardia et al. 1979), on which new locations can be created, typically by mixtures of nodes coordinates. But no such underlying continuum exists in general for non-Euclidean distances d, where the only candidates of locational optima of any kind are restricted *to the n nodes themselves.*

Transformations of distances, as defined by the penalty and potential functions $F(d)$ and $G(d)$, are in that respect relevant to pure and applied geometry, besides their ability to assess nodes centrality and accessibility. Interestingly enough:

- the shortest-path distance d^{sp} is not Euclidean, but general theorems (Joly and Le Calvé 1986) guarantee the existence of a threshold $\alpha_0 \in (0,1)$ such that $F(d^{sp}) = (d^{sp})^\alpha$ *is indeed* an Euclidean distance, for any $\alpha \in [0, \alpha_0]$ (Figures 8 and 9)

- short-range penalties of the form $F_\lambda(d^e) = 1 - \exp[-\lambda(d^e)^2]$, with $\lambda > 0$, known as *Gaussian radial basis kernel* in Machine Learning, transforms an Euclidean distance d^e into a squared Euclidean distance. Such is also the case of mixtures (over

λ) of penalties $F_\lambda(d^e)$. As a matter of fact, the latter class of so-called *Schoenberg transformations* does constitute the most general class of Euclidean-preserving transformations (e.g. Bavaud 2011).

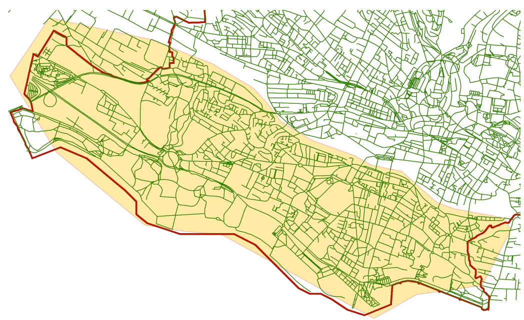

Figure 8: Pedestrian network, district South. Transforming the shortest-paths distances as $(d^{sp})^\alpha$ produces Euclidean distances for $0 \leq \alpha \leq 0.37$

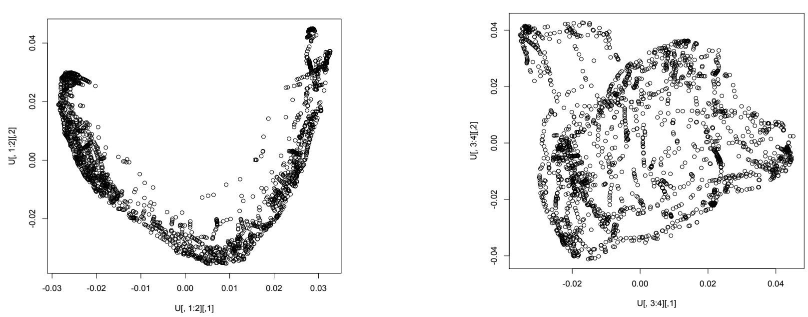

Figure 9: Exact multidimensional scaling (MDS) nodes reconstruction of the pedestrian network of district South, from the Euclidean distances $d^e := (d^{sp})^{0.37}$. Nodes are embedded into a higher-dimensional Euclidean space, thus disrupting the planarity of the network. The first four factorial dimensions depicted above express only 32% of the total inertia $\frac{1}{2}\sum_{ij} f_i^A f_j^A (d_{ij}^e)^2$.

4 Clustering

Another stimulating development considers the construction of new *clustering algorithms* aimed at partitioning the city into m regions, optimal from the point of view of pedestrians in that they minimise the average spatial trip penalty to facilities (schools, shops, services) located (or not) at the centroids of the regions. Clustering the network consists in attributing each of the n nodes to one group $g = 1, \ldots, m$, where the number of groups m is here hold as fixed. In general, the clustering attempts to minimise a given *objective function* such as the total cost $C(Z, L)$ in (2), or to maximise the *total accessibility*

$$A(Z, L) := \sum_{i=1}^{n} \sum_{g=1}^{m} f_i z_{ig} G(d_{iL(g)}) \tag{4}$$

where $G(d)$ is a distance-deterrence function.

If $F(d)$ is an Euclidean distance, the optimal locations $L(g)$ can be freely chosen in the underlying continuous embedding space (section 3.2), as in the K-means algorithm. In the general case, $F(d)$ is not Euclidean and the optimal location candidates $L(g)$ are restricted within the set of existing nodes, as assumed here.

In any case, the resulting clusterings are *hard*, that is node i is attributed to the (supposedly unique) group $g = \arg\min_{h=1}^{m} F(d_{iL(h)})$ for the total cost clustering, or $g = \arg\max_{h=1}^{m} G(d_{iL(h)})$ for the total accessibility clustering. In particular, the memberships z_{ig} take on binary 0/1 values only.

Soft clusterings, allowing multiple nodes-groups attributions, may also be produced by minimizing an "entropy-augmented cost" of the form

$$\min_{Z,L} \quad C[Z, L] + TI[Z] \qquad = \qquad \min_{Z} \quad C[Z, L(Z)] + TI[Z]$$

where the additional entropy term $I[Z]$ is the *nodes-group mutual information*, and $T > 0$ is a free parameter, favoring multiple memberships, interpretable as a *temperature* (Rose 1998; Bavaud 2010). Soft clusterings may be preferred because of their better aptitude to adequately characterise nodes located at the border of two groups, or for improving iterative clustering algorithms by "simulated annealing" techniques (Rose 1998).

In the soft, cost-driven, non-Euclidean case, the optimal membership $Z = (z_{ig})$ and locations $L(g)$ can be shown (e.g. Bavaud 2010) to obey the conditions (yielding to an obvious iterative solving scheme):

$$\rho_g = \sum_{i=1}^{n} f_i z_{ig} \qquad L(g) = \arg\min_{i=1}^{n} \sum_j \frac{f_j z_{jg}}{\rho_g} F(d_{ij}) \qquad z_{ig} = \frac{\rho_g \exp(-\beta F(d_{iL(g)}))}{\sum_{h=1}^{m} \rho_h \exp(-\beta F(d_{iL(h)}))} \tag{5}$$

Here $\beta := 1/T > 0$ is the inverse temperature controlling for the sharpness of the groups.

In the limit $\beta \to \infty$, one recovers the usual *hard clustering iterating procedure*, where the group $g = g[i]$ to which i belongs, and its location $L(g)$ are determined by

$$g[i] = \arg\min_{h=1}^{m} F(d_{iL(h)}) \qquad\qquad L(g) = \arg\min_{i=1}^{n} \sum_{j \in g} f_j F(d_{ij}) = \arg\min_{i=1}^{n} e_i^g \qquad (6)$$

Another family of spatial clusterings arises from *constrained facilites problems*, where locations $L^*(g)$ and weights ρ_g^* are fixed, as for schools, each endowed with a relative capacity ρ_g^* with $\sum_g \rho_g^* = 1$. The constrained total cost problem

$$\min_Z C(Z, L^*) = \min_Z \sum_{i=1}^{n} \sum_{g=1}^{m} f_i z_{ig} F(d_{iL^*(g)}) \qquad \text{with} \sum_i f_i z_{ig} \stackrel{!}{=} \rho_g^*$$

is nothing but the famous *optimal transportation problem* of operations research. Interestingly enough, in the squared Euclidean case $F(d^e) = (d^e)^2$, the objective can be shown to reduce to

$$C(Z, L^*) = \Delta_W[Z] + \sum_g \rho_g^* (d^e_{L(g)L^*(g)})^2 \qquad (7)$$

where $\Delta_W[Z]$ is the *within-cluster dispersion* (Bavaud 2010), to be minimised in Ward clustering method, and $d^e_{L(g)L^*(g)}$ is the distance between the gravity center of the group $L(g)[Z]$ and the real location $L^*(g)$.

A soft clustering scheme for constrained facilites problems can be constructed following the same lines as the unconstrained case, and results in memberships - compare with (5):

$$z_{ig} = \frac{\rho_g^* \epsilon_g \exp(-\beta F(d_{iL^*(g)}))}{\sum_h \rho_h^* \epsilon_h \exp(-\beta F(d_{iL^*(h)}))} \qquad \text{where} \quad \frac{1}{\epsilon_g} = \sum_i \frac{f_i \exp(-\beta F(d_{iL^*(g)}))}{\sum_h \rho_h^* \epsilon_h \exp(-\beta F(d_{iL^*(h)}))} \qquad (8)$$

Here $\epsilon_g > 0$ is a parameter controlling for the weights constraints ρ_g^*. Its value can be iteratively determined by the second identity in (8).

4.1 Illustrations

Let us restrict to the most straightforward (zero-temperature, unconstained) clustering algorithms, aimed at minimising the total cost (2), and consisting of the iterative scheme contained in (6). The resulting hard partitionings, initiated by the same set of $g = 1, \ldots, m = 4$ of initial centroid locations $L_0(g)$, are depicted in Figure 10.

Shortest-path and Euclidean clusters behave similarly in this example, and demonstrate a noticeable contrast between the weighted, versus the unweighted version. The latter is virtually identical to the (weighted or unweighted) squared Euclidean case.

Figure 10: Total cost-based clustering ot the North-West district, unconstrained, hard, weighted (left) and unweighted (right), on $m = 4$ groups. Top: shortest-path distance $F(d^{sp}) = d^{sp}$. Middle: Euclidean distance $F(d^e) = d^e$. Bottom: squared Euclidean distance $F(d^e) = (d^e)^2$.

5 Conclusion

Pedestrian networks are transportation networks, intimately coupled to other transportation networks. Sinuosity, eccentricity and accessibility indices can be defined for any undirected network, weighted or not, endowed with a shortest-path and an Euclidean distance. This paper has attempted to demonstrate how those indices, originally rooted in quantitative geography, are equally relevant in classical operations research and data analysis. In particular, a broad family of nodes clusterings (constrained or not, hard or soft, Euclidean-embedded or not) can be unified in a common framework, based on those quantites. Among other results, Euclidean transformations and representations of the network, as well as identities (7) and (8) appear as original. Future work will carry on the application and development of the descriptive tools and clustering algorithms in an applied spatial planning perspective, such as the assessment, creation or modification of pedestrian short-cuts as well as school locations and capacities in the city of Lausanne.

References

[1] F. Bavaud (2010) *Aggregation invariance in general clustering approaches*, Advances in Data Analysis and Classification 3, pp. 205–225

[2] F. Bavaud (2011) *On the Schoenberg transformations in data analysis: theory and illustrations*, Journal of Classification 28, pp. 297–314

[3] T. Emmanouilidis (2012) *Improving the connectivity of a pedestrian network through Euclidean and shortest path distances comparison*, contributed paper to AGILE 2012 (Avignon)

[4] S. Joly, G. Le Calvé (1986) *Etude des puissances d'une distance*, Statistique et Analyse des Données 11, 29–50

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

[6] L. Kaufman, P. Rousseeuw (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley

[7] S. L. Hakimi (1965) *Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems*, Operations Research 13, 462–475

[8] K. Rose (1998) *Deterministic Annealing for clustering, compression, classification, regression, and related optimization problems*, Proceedings of the IEEE 86, 2210–2239