REVUE DE STATISTIQUE APPLIQUÉE

MAURICE ROUX

Notes sur l'arbre de longueur minima

Revue de statistique appliquée, tome 23, n° 2 (1975), p. 29-35

<http://www.numdam.org/item?id=RSA_1975__23_2_29_0>

NOTES SUR L'ARBRE DE LONGUEUR MINIMA [1]

Maurice ROUX
Laboratoire de statistique - Université Paris 6

Etant donné un ensemble fini S à n éléments qu'on peut supposer représentatifs d'objets naturels, et leurs distances mutuelles, sensées quantifier les dissemblances entre ces objets, l'arbre de longueur minima (en abrégé : a.l.m.) sur S constitue une schématisation des liaisons entre les éléments de S, en ce qu'il fournit une ossature du graphe complet valué A ayant pour arêtes les $n(n-1)/2$ segments distincts d'extrêmités les points de S et de longueurs les distances données entre ces extrêmités.

Toujours en termes de théorie des graphes (Cf. Berge, 1963) il s'agit d'un arbre sur S (i.e. un graphe connexe et sans cycle), ayant les points de S pour sommets, tel que la somme des longueurs de ses arêtes soit minimale. On démontre que, si toutes les distances interindividuelles sur S sont distinctes, ce que nous supposerons désormais, alors cet arbre est unique. Kruskal ainsi que Prim ont publié un algorithme très simple pour construire cet arbre. L'objet de ces notes est d'exposer un autre algorithme à but identique, mais évitant le tri des distances. Une telle opération est en effet couteuse en temps de calcul et encombrante pour la mémoire de notre auxiliaire indispensable : l'ordinateur.

Nous poserons d'abord quelques définitions qui nous permettront de fixer nos notations et de démontrer les propriétés susceptibles de justifier notre algorithme que nous exposerons ensuite.

Le principe de cet algorithme avait été énoncé parmi six autres par P. Rosensthiel (1966), sous l'appellation Kruskal-I dans le désordre ; cet auteur soulignait d'ailleurs judicieusement que c'est le mode de lecture des données (c. à d. des arêtes et de leurs longueurs) qui donne l'avantage à tel ou tel de ces algorithmes.

I. – DEFINITIONS, NOTATIONS ET PREMIERES PROPRIETES DE l'a.l.m.

Dans tout ce qui suit on appelle S l'ensemble fini des objets étudiés, ou ensemble des sommets, et d la distance sur S. On pose Cardinal (S) = Card (S) = n. On désigne par [p] l'ensemble des entiers naturels de 0 à p inclusivement, et par]p] l'ensemble des entiers de 1 à p inclusivement.

---

(1) Article remis le 12/6/73, révisé le 15/11/74.

Définition 1 (Arête) : on appelle arête un sous-ensemble de S à deux éléments, ses extrêmités. On appelle A l'ensemble de toutes les arêtes de S ; Card (A) = $n(n-1)/2$.

Définition 2 (Longueur d'une arête) : on appelle longueur d'une arête $a = \{s, s'\}$ la valeur $d(a) = d(s, s')$.

Définition 3 (Polygone) : on appelle polygone G un ensemble d'arêtes, c. à. d. un sous-ensemble de A. Le support de G sera la réunion des arêtes de G, c. à. d. l'ensemble des sommets de S qui sont extrêmités d'au moins une arête de G. On pourra parler de polygone G *sur* S ou *dans* S, suivant que le support de G est S ou un sous-ensemble de S distinct de S lui-même. A, qui est l'ensemble de toutes les arêtes possibles sur S est aussi appelé polygone complet sur S. Pour Berge un graphe est la conjonction d'un polygone et de son support.

Définition 4 (Adjacent) : Deux arêtes sont dites adjacentes si elles ont une extrêmité commune. Soit $R \subset S$ et G un polygone dans (ou sur) S. Un élément $s \in S$ est dit adjacent à R si $s \in R$ et si s est extrêmité d'une arête de G dont l'autre extrêmité est élément de R. En particulier, si R est réduit à un seul élément r, on dira que r et s sont adjacents dans G si $\{r, s\} \in G$.

Définition 5 (Degré) : le degré d'un sommet s, élément du support d'un polygone G sur ou dans S est le nombre d'arêtes de G qui contiennent s. Un sommet est dit pendant s'il est de degré égal à un.

Définition 6 (Chaîne) : une chaîne C est un polygone dans S dont les sommets qui forment son support peuvent être ordonnés en une suite $(s_0, s_1, \ldots, s_p)$ telle que :

1) $\forall i \in]p[\quad \{s_{i-1}, s_i\} \in C$

2) $\forall i, j \in [p] \; i \neq j \Rightarrow s_i \neq s_j$

On dit que C joint s_0 à s_p ou qu'elle relie ces deux éléments ; on dit aussi que s_0 et s_p sont les extrêmités de C.

Un cycle est une chaîne où la condition 2 ci-dessus vaut pour tous les points de son support exceptés s_0 et s_p qui sont confondus.

Définition 7 (Connexe) : un polygone G de S est dit connexe si, pour tout s et tout s' du support de G il existe une chaîne incluse dans G dont les extrêmités sont s et s'.

Définition 8 (Arbre) : un arbre est un polygone connexe qui ne contient pas de boucle.

On démontre les théorèmes suivants (Cf. Berge, 1963) :

Théorème 1 : soit G un polygone sur S, Card (S) = n ; les propriétés suivantes sont équivalents :

1) G est connexe et sans cycle.

2) G est sans cycle et possède n − 1 arêtes.

3) G est connexe et possède n − 1 arêtes.

4) G est sans cycle et en ajoutant une arête entre deux sommets non-adjacents on crée un cycle (et un seul).

5) G est connexe et en supprimant une arête quelconque il n'est plus connexe.

6) Tout couple de sommets est relié par une chaîne et une seule.

Théorème 2 : un arbre possède au moins deux sommets pendants

Théorème 3 : si p est le cardinal du support d'un arbre T alors Card (T) = p − 1 (i.e. T possède p − 1 arêtes).

Définition 9 (longueur et largeur) : on appelle longueur d'un polygone la somme des longueurs de toutes ses arêtes ; sa largeur sera la longueur de son arête la plus longue.

On remarquera que la largeur (en arbrégé : larg.) d'un polygone est en général inférieure au diamètre de son support. Ces deux valeurs coïncident si le polygone contient l'arête la plus longue ayant pour extrêmités des points de ce support. On remarquera, en outre, que, pour le polygone réduit à une arête, longueur et largeur sont identiques.

Si G et G′ sont deux polygones dans S, alors :

1) $\text{Larg}(G \cup G') = \max(\text{Larg}(G), \text{Larg}(G'))$

2) si $G \cap G' = \phi$, on a $\text{Longueur}(G \cup G') = \text{Longueur}(G) + \text{Longueur}(G')$.

II. – AUTRES PROPRIETES DE l'a.l.m.

Nous donnons sous ce titre trois propositions permettant de valider notre construction de l'a.l.m. Les deux premières sont connues (Cf. par exemple Kalaba ou Berge), tandis que la troisième est démontrée dans le cadre axiomatique des matroïdes (Rosensthiel, in Journal of combinatorial theory, sous presse).

Proposition 1 : soit T a.l.m. sur S. Alors toute partie connexe T′ de T est a.l.m. sur son support.

S'il n'en était pas ainsi il existerait T″ a.l.m. sur Support (T′) et dans ce cas l'arbre (T − T′) ∪ T″ serait de longueur strictement inférieure à Longueur (T) ce qui est impossible.

Proposition 3 : supposons que le polygone réduit à l'arête $\{s, s'\}$ représente la chaîne de A de largeur minima (nous dirons désormais : la chaîne minimale) joignant s à s′, alors $\{s, s'\}$ est élément de l'a.l.m. T sur S (il existe au moins une telle arête : l'arête de A de longueur minima).

Supposons que cette proposition soit fausse ; il existe alors dans T une chaîne C de support $\{c_0, c_1, \ldots, c_p\}$ avec $c_0 = s$ et $c_p = s'$. Soit $\{c_i, c_{i+1}\}$ le plus long maillon de cette chaîne : $d(c_i, c_{i+1}) > d(s, s')$ par hypothèse ; on obtiendrait donc un arbre de longueur strictement plus petite en remplaçant l'arête $\{c_i, c_{i+1}\}$ par $\{s, s'\}$. On vérifie en effet que le polygone ainsi obtenu est sans cycle et connexe (la suppression de $\{c_i, c_{i+1}\}$ "déconnecte" le polygone, s étant dans l'une des parties connexes, s′ dans l'autre ; l'adjonction de $\{s, s'\}$ redonne un polygone connexe sans introduire de cycle).

Lemme : soit T a.l.m. sur S et $s \in S$ un sommet pendant ; alors pour tout t élément de S différant de s, et quelle que soit l'arête a de C_{st}, chaîne de T reliant s à t, on a $d(s, t) \geqslant d(a)$.

Supposons qu'il existe $a \in C_{st}$ de longueur strictement supérieure à $d(s, t)$. Supprimons a de T ; on obtient ainsi deux parties connexes, l'une contenant s l'autre t. Adjoignons à ce polygone l'arête $\{s, t\}$, on reconstitue de cette façon un arbre (Théorème 1, no 6) de longueur strictement plus petite que celle de T, ce qui est impossible.

Proposition 3 : (Réciproque de la proposition 2) si $\{r, s\}$ est élément de l'a.l.m. T sur S, alors $\{r, s\}$ est la chaîne de A ayant la largeur la plus petite et joignant r à s.

Nous allons démontrer que $\{r, s\}$ est minimale et comme nous avons supposé, depuis le début, que toutes les distances sur S sont distinctes la proposition en résultera. Cette proposition est évidemment vraie si Card (S) = 2. On raisonne donc par récurrence sur le cardinal de S_p support de l'arbre $T_p \subset T$ contenant $\{r, s\}$; on pose $S_p = \{r, s, t_1, \ldots, t_p\}$, $S_{p+1} = S_p \cup \{t_{p+1}\}$ où t_{p+1} est un élément de S adjacent à S_p dans T ; l'hypothèse de récurrence est que la proposition 3 est vraie sur tout sous-ensemble de S_p, y compris S_p lui-même, on veut démontrer qu'elle est encore vraie pour S_{p+1}. Supposons que cela ne soit pas ; ceci signifie que l'apparition de t_{p+1}, que nous noterons t dans la suite, fait qu'il existe une chaîne minimale, C', passant par t et joignant r et s. Soit $a_r = \{r', t\}$ et $a_s = \{s', t\}$ les deux arêtes de cette chaîne issues de t Posons que $T_p^{(t)}$ est la partie connexe de $T - \{r, s\}$ contenant r. On suppose que $r' \in$ Support $(T_p^{(r)})$ et $s' \in$ Support $(T_p^{(s)})$; s'il n'en était pas ainsi (Cf. figures) alors un autre élément de la chaîne C' (différent de a_r et de a_s) ferait le lien entre $T_p^{(r)}$ et $T_p^{(s)}$, dont la longueur serait supérieure à celle de $\{r, s\}$ et C' ne serait pas minimale.

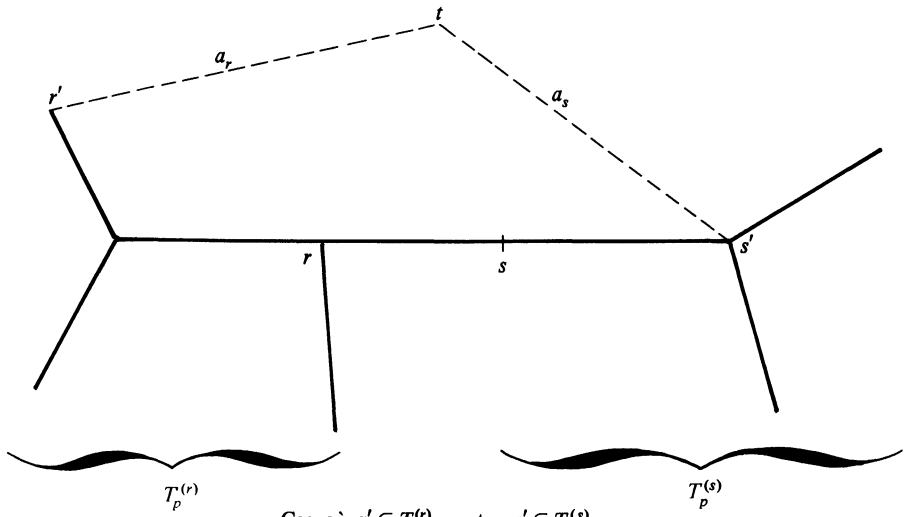
Dans ces conditions appelons $C_{rr'}$ (resp. $C_{ss'}$) la chaîne de $T_p^{(r)}$ (resp. $T_p^{(s)}$) joignant r à r' (resp. s à s') et $C'_{rr'}$ la partie de C' joignant r à r' ($C'_{ss'} \subset C'$ joint s à s') ; dans T_p l'hypothèse de récurrence s'applique, donc $C_{rr'}$ est minimale, entre r et r', sur s_p, de même $C_{ss'}$ est minimale, entre s et s' sur S_p. Il en résulte que $\text{Larg}(C_{rr'}) \leqslant \text{Larg}(C'_{rr'})$ et $\text{Larg}(C_{ss'}) \leqslant \text{Larg}(C'_{ss'})$. On aura donc encore une chaîne minimale en remplaçant C' par la chaîne :

$$C = C_{rr'} \cup a_r \cup a_s \cup C_{ss'}$$

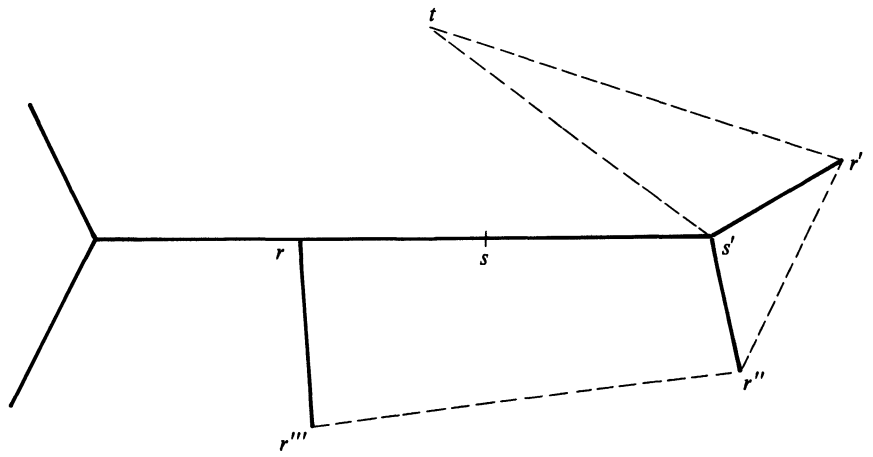
dont chacune des parties est de largeur plus petite que celle des parties homologues de C'.

Or t est sommet pendant de $T_{p+1} \subset T$ de support S_{p+1} donc l'une ou l'autre des deux arêtes a_r et a_s (ou les deux) n'appartient pas à l'a.l.m. et d'après le lemme l'une ou l'autre est plus longue que $\{r, s\}$. Il s'ensuit que $\text{Larg}(C) \geqslant \text{Larg}(\{r, s\})$ et que ni C ni C' ne sont minimales sur S_{p+1} ce qui achève la démonstration.

Algorithme proposé : notre algorithme procède par modificiation progressive d'un arbre initial T_1 arbitraire. On peut choisir celui-ci simplement de la façon suivante (mais tout autre choix est possible) : chacune des $n - 1$ arêtes est formée d'un élément de S, de numéro compris entre 2 et n, l'autre extrêmité étant le numéro 1. On enregistre soigneusement les longueurs de ces arêtes.



Cas où $r' \in T_p^{(r)}$ et $s' \in T_p^{(s)}$



Cas où r' et s' sont éléments de $T_p^{(s)}$

On considère successivement tous les points i de S autres que le premier et le dernier. Pour chacun d'eux on examine les distances d (i , j) avec les autres points j de S non encore considérés ; si la donnée est un tableau rectangulaire de description objets x variables, on peut calculer ces distances à ce moment précis. On compare alors d (i , j) à la largeur de la chaîne C_{ij} joignant i à j dans l'arbre actuel T_k ; si celle-ci est plus petite que celle-là alors $T_{k+1} = T_k$, sinon on enlève de T_k la plus longue arête de C_{ij} que l'on remplace par $\{i , j\}$ pour obtenir T_{k+1}. Lorsque la dernière arête a été envisagée (arête de rang $(n-1)$ $(n-2)/2 = n(n-1)/2 - (n-1)$, car nous ne décomptons pas les $n-1$ arêtes de T_1) l'arbre obtenu est l'a.l.m. cherché.

En effet, supposons qu'on obtienne une arête n'appartenant pas à l'a.l.m. alors celui-ci, pour être connexe, contient un segment qui n'est pas dans l'arbre obtenu. Mais cette arête a été envisagée au cours de l'algorithme, comme toutes

les autres, et puisqu'elle n'a pas été conservée c'est qu'on a trouvé une chaîne de largeur strictement plus petite que la chaîne réduite à cette arête et joignant ses extrêmités, ce qui contredit la proposition 3.

L'intérêt de cette technique nous paraît être de pouvoir traiter, en un temps de calcul très intéressant de gros ensembles de données. Une fois le problème ainsi débroussaillé on peut, par exemple, faire des analyses factorielles sur des parties de l'a.l.m. ou sur les individus jugés représentatifs, les autres étant placés en "éléments supplémentaires" (Cf. Benzécri, 1973), pour affiner les résultats. Flament s'est intéressé à l'étude des chaînes de l'a.l.m., en particulier à celles qui ont un cardinal maximal. Pour notre part nous avons pu constater sur des exemples de phytosociologie (Cf. Benzecri [Alpes II] TIC no 3) qu'une telle chaîne est fréquemment unique et correspond grossièrement au premier axe d'une analyse factorielle des correspondances.

Bien entendu, lorsque nous avons rédigé ces notes, nous n'ignorions pas les travaux de Roger et Carpenter sur la construction cumulative de l'a.l.m. ; néanmoins comme nous avions mis au point notre méthode de façon indépendante et qu'elle est radicalement différente, bien que présentant vraisemblablement les mêmes avantages, nous avons cru utile de la publier d'autant plus que nous proposons un programme de calcul accompagné d'un sous-programme de tracé de l'arbre sur imprimante rapide, ce qui à notre connaissance est original. Ce sous programme est du à C. Désarménien de l'I.R.I.A. (Domaine de Voluceau, 78. – Rocquencourt).

BIBLIOGRAPHIE

BENZECRI J.P. et coll. – Leçons sur l'analyse des données. Dunod, Paris, 1973.

BERGE C. – Théorie des graphes et ses applications. Dunod, Paris, 1963.

BERGE C. – Graphes et hypergraphes. Dunod, Paris, 1970.

DREYFUS S.E. – An appraisal of some shortest path algorithms. *Operations Research* Vol. 17, no 3, pp. 395-412, 1969.

FLAMENT C. – Théorie des graphes et structures sociales. Gauthiers-Villars, Paris, 1965.

GOWER J.C. and ROSS G.J.S. – Minimum spanning trees and single linkage cluster analysis. *Applied Statistics,* 18, pp. 54-64, 1969.

KRUSKAL J.B. – On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. Amer. Math. Soc.* 7, pp. 48-50, 1956.

PRIM R.C. – Shortest connexion network and some generalizations. *Bell Syst. Techn. J.* pp. 1389-1401, 1957.

ROGER J.H. and CARPENTER R.G. – The cumulative construction of minimum spanning trees. *Applied Statistics,* vol. 20, no 2, pp. 192-194, 1971.

ZAHN C.T. – Graph-theoretical methods for detecting and describing gestalt clusters *IEEE Trans. on computers,* vol. 20, no 1, 1971.

ROSENSTHIEL P. – L'arbre minimum d'un graphe. In "Théorie des graphes", proceedings of the international symposium, Rome, 1966. Ed. Dunod, Paris 1967.

KALABA R. – Graph theory and automatic control, in "Applied combinatorial mathematics", E.F. Beckenbach ed. pp. 237-252, Wiley, 1969.