GéoPeuple

Rapport numéro	L2.3-1-2
Titre	Appariement Cassini - RGE (appariement des réseaux)
Rédigé par	Benoit Costes (COGIT/IGN)
État (en cours / final)	Final
Relu par	Julien Perret (COGIT/IGN), Éric Grosso (COGIT/IGN)
Date	Mai 2013

Nous présentons dans ce rapport la méthode mise au point dans le cadre du projet GéoPeuple pour apparier le réseau hydrographique issue des cartes de Cassini avec la BDCarto de l'IGN, une base de données topographique actuelle de référence. La spécificité de l'approche proposée permet de prendre en compte les difficultés liées à l'appariement des données anciennes, détaillées dans [Costes et al., 2012 ; Costes, 2012] : une différence de temporalité majeure entre les données (environ 250 ans) impliquant des transformations des entités cartographiées, une différence de niveau de détail et de représentation, ainsi que des imperfections dues à l'imprécision des fonds de cartes (décalages de positions importants entre objets homologues, divergences et incomplétude de la toponymie, incertitude de la localisation relatives des objets, etc.).
Afin de tester l'efficacité de l'approche proposée, nous avons prélablement réalisé un appariement manuel du réseau hydrographique issu des cartes de Cassini sur les trois zones considérées dans le projet GéoPeuple (Reims, Grenoble et Saint-Malo).

0.1 L'appariement de géométries linéaire

0.1.1 Mesures d'évaluation de l'appariement

Afin d'évaluer la qualité d'une méthode d'appariement, deux mesures sont couramment utilisées dans le domaine de l'information géographique : la précision et le rappel. Nous appelons classiquement :
– vrais positifs les liens d'appariements correctement établis par le processus automatique (vp)
– faux positifs les liens établis par erreur (fp)
– vrais négatifs les objets correctement non appariés (vn)
– faux négatifs les objets non appariés par erreur par le processus automatique (fp)
Dans la logique d'évaluation d'un appariement de données géographiques, il convient de s'intéresser aussi bien aux liens d'appariement qu'aux objets non appariés, notamment dans un contexte d'appariement de données anciennes. La *précision* des liens d'appariement P_{app} (respectivement des objets non appariés P_{nap}) est le rapport entre le nombre de vrais positifs (respectivement de vrais négatifs) et la somme des vrais positifs et des faux positifs (respectivement des vrais négatifs et des faux négatifs). Le *rappel* des liens d'appariement R_{app} (respectivement des objets non appariés R_{nap}) est le rapport entre le nombre de vrais positifs (respectivement de vrais négatifs) et le nombre de liens attendus,

c'est-à-dire établis manuellement. Notons N_{app} le nombre de liens d'appariement attendus, et N_{nap} le nombre d'objets non appariés attendu. Alors :

$$P_{app} = \frac{vp}{vp + fp} \tag{1}$$

$$R_{app} = \frac{vp}{N_{app}} \tag{2}$$

$$P_{nap} = \frac{vn}{vn + fn} \tag{3}$$

$$R_{nap} = \frac{vn}{N_{nap}} \tag{4}$$

À notre avis, une bonne mesure d'évaluation ne doit être ni trop optimiste, ni trop pessimiste. Nous utilisons une mesure qui combine équitablement la précision et le rappel, appelée *F-score* et définie comme suit :

$$F\text{-}score = 2 * \frac{precision * rappel}{precision + rappel} \tag{5}$$

0.1.2 Positionnement par rapport à l'existant

De nombreuses approches d'appariement de réseaux issues de la littérature considèrent le réseau comme un graphe et se basent sur des analyses géométriques (distances, angles, plus courts chemins) [Zhang et al., 2005] et topologiques (degré des nœuds, relations de connectivité, etc.), déduisant les appariements des arcs d'un appariement préalable des nœuds du réseau [Mustière and Devogele, 2008 ; Lüscher et al., 2007]. Les importants décalages géométriques des réseaux à apparier dans notre étude rendent ces approches peu efficaces dans de nombreuses situations (c.f figure 1, c.f tableau 1). Ces divergences géométriques peuvent trouver leur origine dans le fait que les représentations cartographiques des objets du monde réel faîtes à l'époque de Cassini étaient fortement subjectives et influencées par les méthodes de saisie de l'époque. Nous remarquons par exemple des bifurcations ou des contournements de cours d'eau au niveau des villes, imputables non pas à la topologie du réseau hydrographique de l'époque, mais à une volonté des cartographes de privilégier l'information religieuse et urbaine (églises) en décalant volontairement le tracé des cours d'eau qui en gênaient le dessin ou l'inscription du toponyme. De plus, en zones montagneuse, la qualité géométrique des représentations des cours d'eau, et leur positionnement relatifs étaient difficiles à retranscrire à cause du relief.

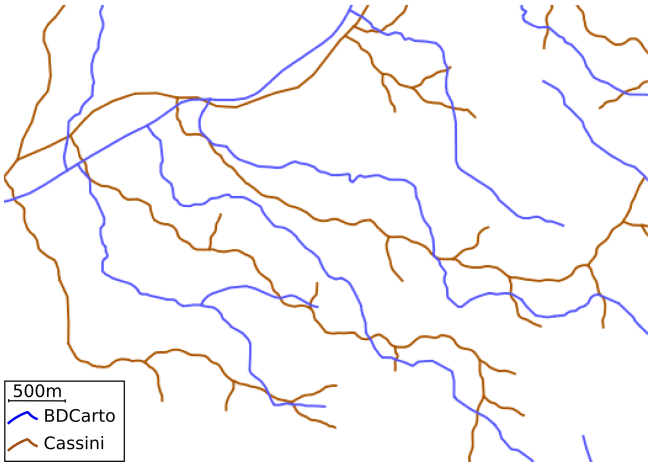


FIGURE 1: Exemple de situation posant problème à une approche d'appariement basée uniquement sur la géométrie et la topologie du réseau

D'autres approches proposent l'utilisation d'une méthode multi-critères prenant en compte à la fois la géométrie, la sémantique, la toponymie, voire également la topologie [Olteanu, 2008 ; Costes et al.,

	liens d'appariement			objets non appariés		
Zones	**Précision**	**Rappel**	**FScore**	**Précision**	**Rappel**	**FScore**
Approche de [Mustière and Devogele, 2008]						
Reims	73%	85%	78%	44%	80%	57%
Grenoble	58%	76%	66%	89%	85%	87%
St-Malo	63%	91%	74%	86%	71%	77%
Approche de [Costes et al., 2012]						
Reims	90%	87%	89%	57%	80%	67%
Grenoble	67%	82%	74%	97%	85%	90%
St-Malo	48%	68%	56%	86%	71%	77%

TABLE 1: Évaluation de l'appariement utilisant les approches de [Mustière and Devogele, 2008] et de [Costes et al., 2012] sur les données d'étude : précisions, rappel et FScore des liens d'appariement (LAP), précision, rappel et FScore des objets non appariés (NAP)

2012], permettant de prendre en compte les imprécisions des données. Cependant, les résultats d'appariement obtenus par ces méthodes ne sont pas entièrement satisfaisants (c.f tableau 1). Nous expliquons ces résultats par l'incomplétude majeure touchant à la fois la sémantique et la toponymie du réseau hydrographique de Cassini. En effet, une grande majorité des cours d'eau ne sont pas nommés (plus de 90% de toponyme "Non renseigné"), ou sont de nature inconnue (environ 85% de nature "indéterminée"). De plus, ces approches, qu'elles soient purement multi-critères [Costes et al., 2012], ou statistiques[Walter and Fritsch, 1999], nécessitent systématiquement des connaissances supplémentaires pour définir par apprentissage les paramètres utiles à l'appariement.

0.2 Propositions

Nous proposons une nouvelle approche multi-critères basée sur un appariement hiérarchique des tronçons, en identifiant les cours d'eau principaux, classés selon une relation d'importance prenant en compte les ramifications du réseau. Son originalité réside également dans la définition d'un critère supplémentaire utilisant la proximité d'objets ponctuels appariés selon [Costes et al., 2012] comme pivot pour l'appariement, en posant l'hypothèse que *la plupart des objets subsistant de nos jours (villes, hameaux, moulins ...) traversés par ou proches des cours d'eau principaux dans le passé le sont également aujourd'hui.*

0.2.1 Enrichissement des réseaux

Dans une première étape, nous rendons les réseaux plus comparables en dégageant les structures naturelles continues, appelées "strokes" [Thomson and Richardson, 1999 ; Jiang et al., 2008], constituées d'une succession de tronçons et représentant réellement les objets hydrographiques : fleuves, rivières, ruisseaux, etc. Le terme de "stroke" vient de l'idée que l'élément peut être dessiné d'un seul trait lisse, doux et sans à-coups [Thomson and Richardson, 1999]("in one stroke" peut signifier "en un seul coup") reprenant le principe de continuité de la Gestalt : "Continuation occurs because the viewer's eye will naturally follow a line or curve".
Si à l'échelle du tronçon, les niveaux de détail et la géométrie des réseaux diffèrent, il sont relativement proches quand on raisonne à l'échelle des strokes.

0.2.1.1 Prétraitement : préparation des réseaux

0.2.1.2 Sélection des strokes

[Thomson and Richardson, 1999 ; Jiang et al., 2008] construisent les strokes sur des critères angulaires aux niveau des nœuds, et attributaires au niveau des tronçons. L'approche proposée par [Touya, 2010]

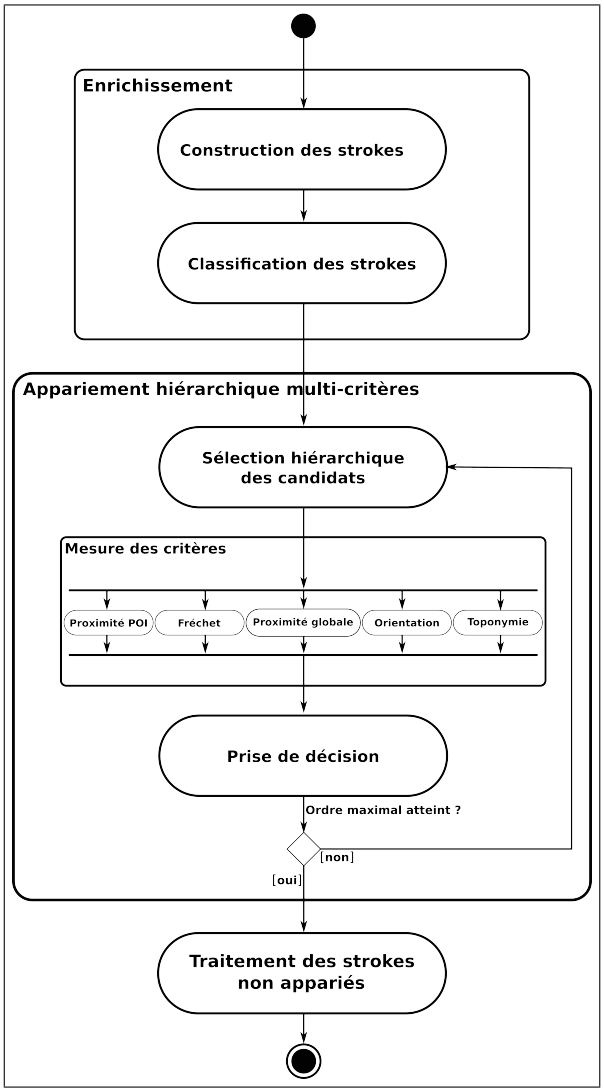
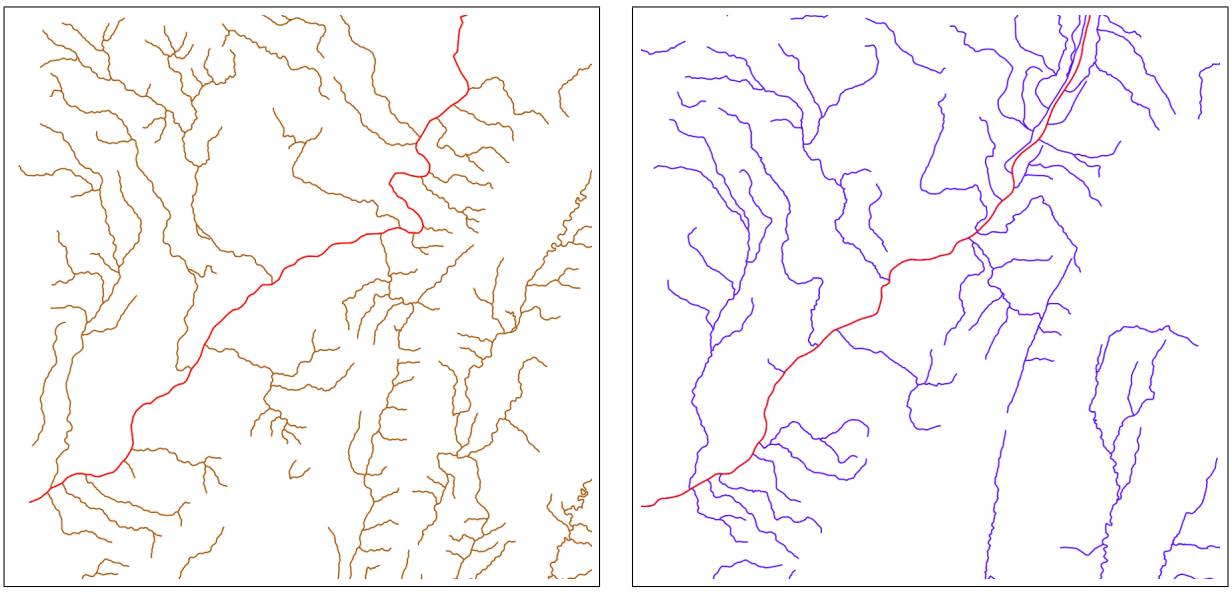


FIGURE 2: Schéma global de notre approche d'appariement hiérarchique multi-critères.

repose également sur des critères de continuité angulaire aux intersections d'arcs, et de continuité attributaire, tout en détectant en amont les structures particulières plus délicates à traiter (méandres, îles simples et complexes, zones d'irrigation, etc.)
Nous adoptons une approche qui étend l'algorithme "every-best-fit" de [Jiang et al., 2008], qui travaille de la façon suivante. Lorsque plusieurs arcs partagent un nœud commun, nous joignons la paire d'arc ayant la meilleur continuité en ce nœud, évaluée de la manière suivante :

- dans un premier temps, nous déterminons si la continuité toponymique est respectée : tous les arcs d'un stroke doivent avoir le même toponyme si au moins un d'entre eux est nommé,
- si la toponymie ne permet pas de trancher, nous évaluons la continuité sémantique : tous les arcs doivent être de la même nature (rivière ; fleuve, etc.) si la nature d'au moins un d'entre eux est renseignée,
- enfin, en dernier recours si ni la continuité toponymique, ni la continuité sémantique ne permettent de choisir, nous évaluons la continuité angulaire, en choisissant la paire d'arc ayant le plus petit angle de déflexion.

Le détail de l'algorithme est donné en annexe.



(a) (b)

FIGURE 3: Exemple de structure naturelle continue (stroke) : la rivière Isère, pour Cassini (a) et la BDCarto (b). Les échelles et emprises sont identiques.

0.2.1.3 Ordre d'un stroke

[Thomson and Brooks, 2000] et [Touya, 2010] utilisent pour généraliser des réseaux hydrographiques une classification de Horton [Horton, 1945], qui attribue à chaque stroke un degré d'importance en fonction de son nombre d'affluents :"tout cours d'eau sans affluent est d'ordre 1, tout cours d'eau ayant un affluent d'ordre n est d'ordre n + 1, et garde cet ordre sur toute sa longueur". Le problème de cette classification est sa forte dépendance au niveau de détail des réseaux considérés. En effet, comme la classification démarre des feuilles, le stroke principal (sans affluent) d'un réseau très détaillé aura un ordre potentiellement très différent du stroke homologue dans une base à grranularité plus faible.
Nous reprenons ce type de classification en initialisant l'algorithme sur les racines du réseau plutôt que sur les feuilles. A l'inverse d'une classification de Horton, nous attribuons donc l'ordre 1 aux strokes les plus importants, qui ne sont affluents d'aucun autre stroke, et construisons cette relation d'importance récursivement : un stroke affluent d'un stroke d'ordre n, est d'ordre $n+1$. La classification ainsi définie permet d'attribuer le même ordre (ordre 1) aux strokes les plus importants dans différentes bases de données, indépendamment du niveau de détail et du nombre d'affluents. L'ordre du stroke reflète bien ainsi son positionnement dans l'arbre des ramifications du réseau hydrographique et donc son importance en terme de structure hydrographique. Le défaut de notre classification est sa sensibilité à l'emprise des données. Nous nous assurons que le découpage du réseau n'insère pas d'artefacts dans la détermination de l'ordre des strokes.
Pour la suite de ce rapport, nous appelons stroke parent d'un stroke d'ordre n le stroke d'ordre $n-1$ dont il est affluent : les strokes issus des ramifications d'un stroke "parent" d'ordre n sont donc d'ordre $n+1$.

0.2.2 Appariement hiérarchique multi-critères

Nous proposons une approche d'appariement basée sur l'utilisation conjointe de différents critères et reposant sur le respect de la hiérarchie dans la structure du réseau des strokes associés aux tronçons à apparier. L'algorithme est itératif, et pour une passe du processus donnée, nous procédons en trois étapes :
– sélection hiérarchique des strokes candidats à l'appariement : construction de couples de candidats

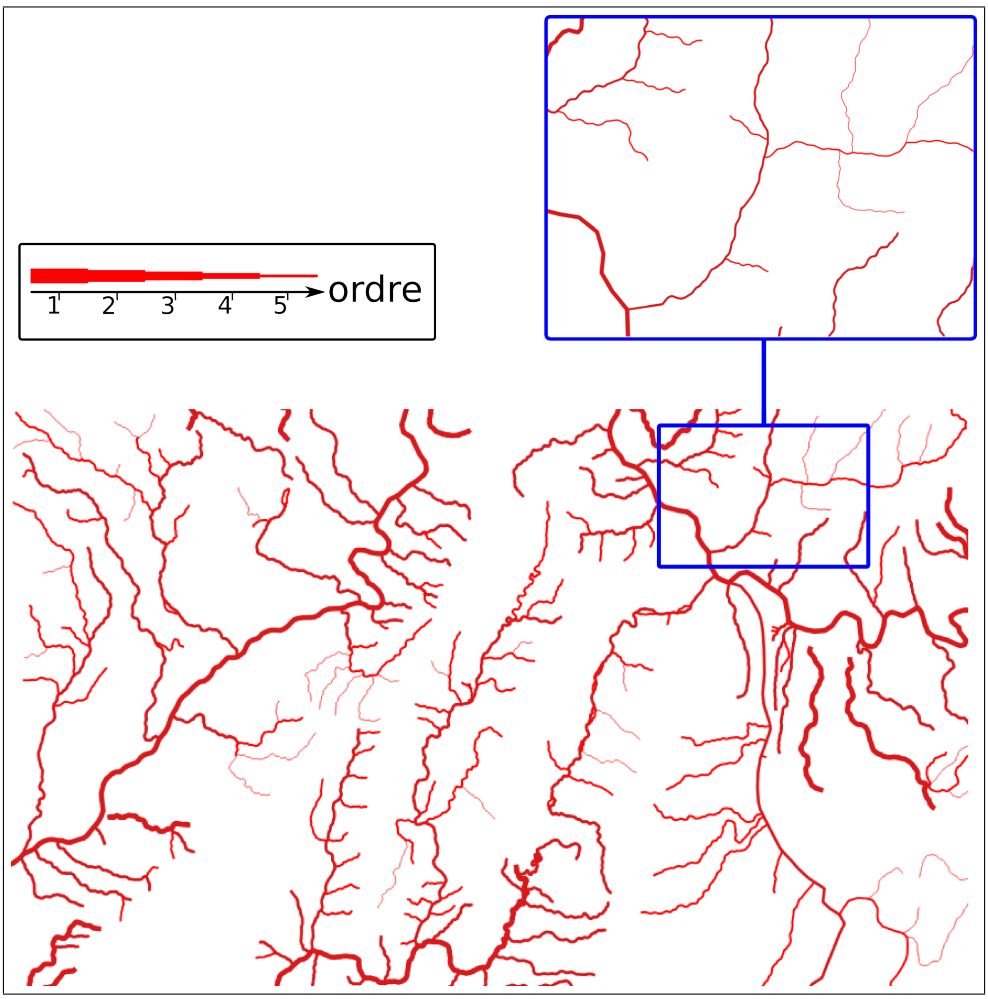


FIGURE 4: Relation d'importance des strokes basée sur la ramification du réseau hydrographique. Plus un stroke est épais, plus son ordre est faible.

$\{s_1, s_2\}$ où *s1* appartient au premier réseau et *s2* au second,
– mesure de chaque critère pour chaque couple de candidats,
– prise de décision
La sélection des couples candidats à l'appariement est ainsi dépendante de la structure hiérarchique des réseaux, et des appariements précédents.

0.2.2.1 Définition des critères utiles à l'appariement

Nous utilisons cinq critères pour l'appariement des strokes.

0.2.2.1.1 Critère de proximité des entités à géométrie ponctuelle Ce critère est exprimé par l'hypothèse posée dans la partie 0.1.2 que nous rappelons : la plupart des objets subsistant de nos jours (villes, hameaux, moulins ...) traversés par ou proches des cours d'eau principaux dans le passé le sont également aujourd'hui.
En effet, dans le contexte du projet GéoPeuple, il est pertinent de s'intéresser aux cours d'eau majeurs, desservant les lieux-dits et potentiellement acteurs des modifications les affectant. Ces lieux-dits, proches ou traversés par un tronçon hydrographique au 18^{me} siècle, le sont toujours à notre époque sauf détournement exceptionnel du cours d'une rivière à proximité d'une grande ville, situation que à laquelle nous n'avons pas été confronté sur les trois zones d'étude choisies pour le projet . Identifier les villes, bourgs, hameaux appariés et proches d'un cours d'eau dans chaque base de données permet de

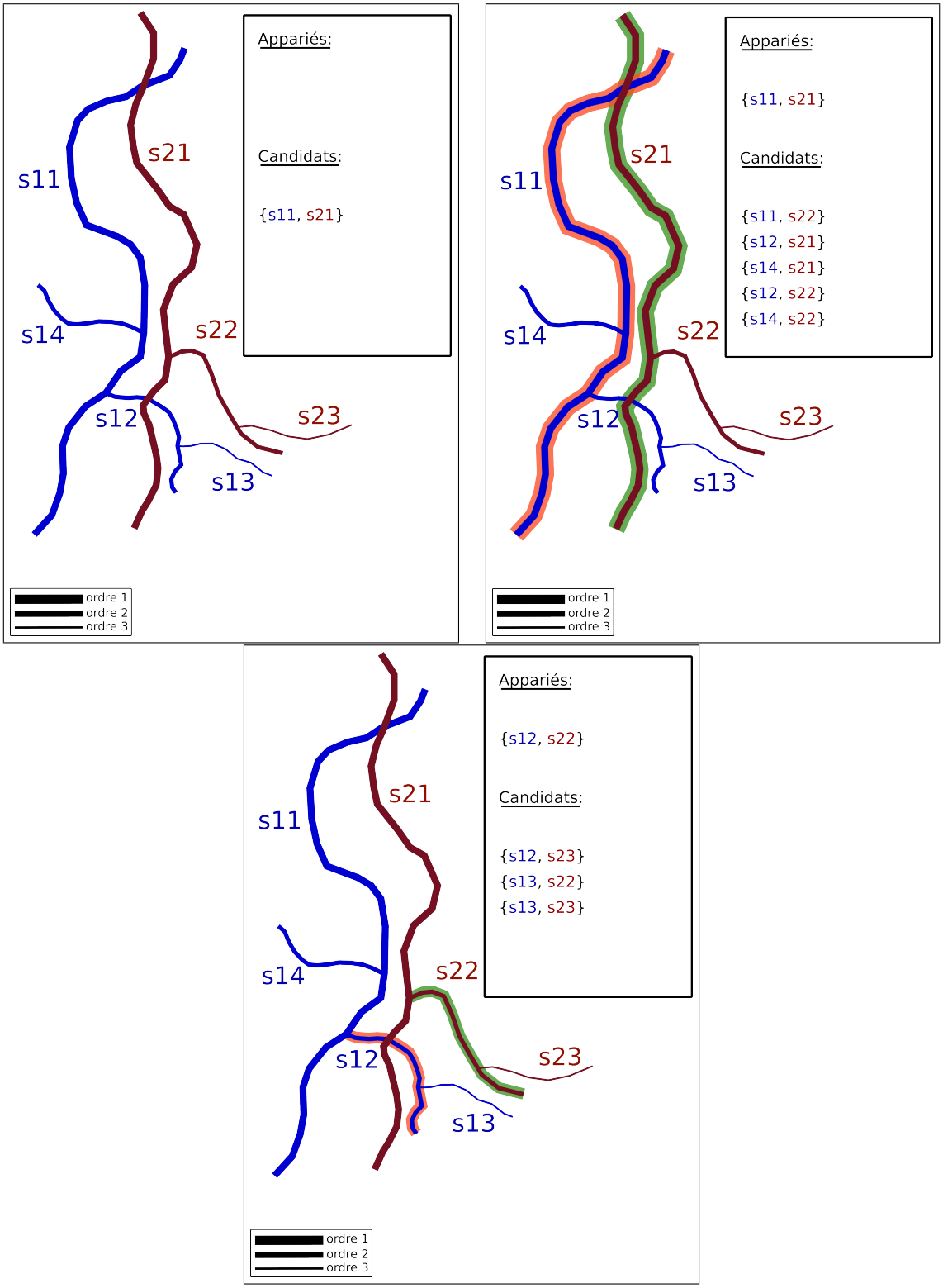


FIGURE 5: Sélection hiérarchique des strokes candidats à l'appariement : les 3 premières passes du processus sur un exemple.

supposer que la correspondance des cours d'eau respectifs est avérée.
Nous utilisons donc pour ce critère les résultats de l'appariement des classes d'objets Cassini à géométrie ponctuelle : le religieux (correspondant en grande partie aux lieux-dits majeurs du 18^{me} siècle), le non religieux (correspondant aux lieux-dits de moindre importance) et les moulins (moulin à eau, à vent, forge, etc.). L'appariement a été effectué avec la BDTopo, une autre base de données de référence de l'IGN. Nous supposons que les liens d'appariement obtenus par la méthode présentée dans [Costes et al., 2012] ont été validés par un post-traitement manuel et sont certains. Nous donnons un exemple de visualisation de cette hypothèse sur la figure 6. Nous constatons que les homologues des entités Cassini proches du cours d'eau du 18^{me} siècle sont peu distants du cours d'eau actuel, la notion de proximité étant définie par une zone tampon, ici matérialisée par des buffers de couleur orange (Cassini) et bleu (BDTopo). Nous mesurons ce critère en calculant deux proportions. Soit s_1 (respectivement

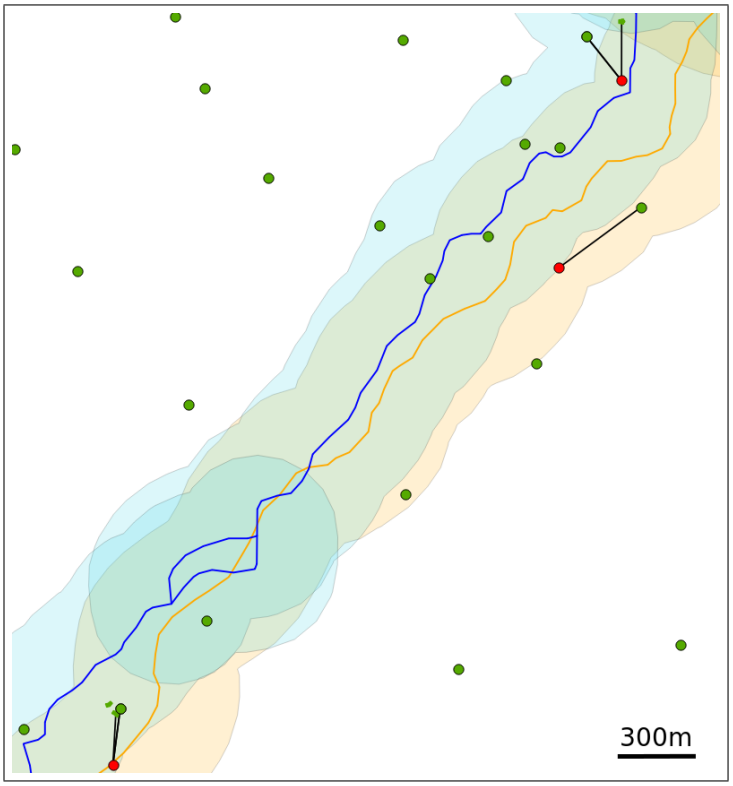


FIGURE 6: Respectivement en orange et en bleu, un bras de rivière issu de la carte de Cassini (Reims) et de la BDTopo ; en rouge et en vert, les entités ponctuelles de Cassini et les objets de la BDTopo. Les liens d'appariement déterminés par la méthode de [Costes et al., 2012] sont en traits noirs.

s_2) le stroke candidat à l'appariement du réseau hydrographique de Cassini (respectivement de la BDCarto). Soit n_1 (respectivement n_2) le nombre d'entités ponctuelles Cassini (respectivement de la BDTopo) appariés proches de s_1 (respectivement de s_2). Enfin, soit $n_{1\to2}$ (respectivement $n_{2\to1}$) le nombre d'objets ponctuel Cassini proches de s_1 et appariés avec un objet ponctuel de la BDTopo proche de s_2 (respectivement le nombre d'objets ponctuels de la BDTopo proches de s_2 et appariés avec un objet ponctuel Cassini proche de s_1).
Le rapport $p_{1\to2} = \frac{n_1}{n_{1\to2}}$ (respectivement $p_{2\to1} = \frac{n_2}{n_{2\to1}}$) représente la proportion d'objets Cassini proches de s_1 et appariés avec un objets de la BDTopo proche de s_2 (respectivement la proportion d'objets de la BDTopo proches de s_2 et appariés avec un objets Cassini proche de s_1).
La valeur mesurée pour ce critère est finalement donnée par le maximum de ces deux proportions : $max(p_{1\to2}, p_{2\to1})$. Plus la valeur mesurée est proche de 1, plus ce critère est favorable à l'appariement des deux candidats.

0.2.2.1.2 Critère de distance géométrique Nous mesurons l'éloignement entre les deux strokes candidats à l'appariement en calculant leur distance de Fréchet discrète (ou coupling frechet distance), une approximation de la distance de Fréchet pour les courbes polygonales (ou polylignes) [Eiter, 1994]. Une courbe polygonale P est une fonction $P : [|0, n|] \rightarrow M$ où $n \in \mathbb{N}$ et M est un espace métrique, telle que $\forall i \in \{0, 1, ..., n-1\}$, la restriction de P à l'intervalle $[|i, i+1|]$ est affine.
Soient $P1$ et $P2$ deux polylignes définies par $P1 : [|0, n|] \rightarrow M, P2 : [|0, m|] \rightarrow M$. Notons $\sigma(P1) = (u_1, ..., u_n) = (P1(0), ..., P1(n))$ et $\sigma(P2) = (v_1, ..., v_m) = (P2(0), ..., P2(m))$. Un *coupling* L entre $P1$ et $P2$ est une suite $(u_{a_1}, v_{b_1}), ..., (u_{a_k}, v_{b_k})$ de paires distinces de $\sigma(P1) * \sigma(P2)$ telle que $a_1 = 1, b_1 = 1, a_k = n, b_k = m$ et $\forall i \in [|1, k-1|], a_{i+1} = a_i$ ou $a_{i+1} = a_i + 1$ et $b_{i+1} = b_i$ ou $b_{i+1} = b_i + 1$. Sa longueur $||L||$ est définie par $||L|| = \max_{i \in [|1,k|]} (d(u_{a_i}, v_{b_i})$ avec d la distance usuelle sur M. La distance de Fréchet discrète entre $P1$ et $P2$ est alors définie par : $dF(P1, P2) = min\{||L||$ avec L un coupling entre P1 et P2$\}$.

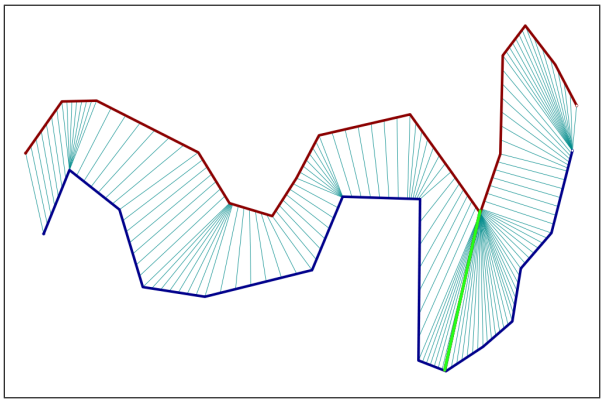


FIGURE 7: Calcul de la distance de Fréchet discrète

La distance de Fréchet permet de rendre compte non seulement de l'écart entre deux courbes, mais aussi de leur différence de forme. Le calcul de la distance de Fréchet discrète repose sur une implémentation par programmation dynamique et s'effectue en temps polynomial.
Une illustration couramment utilisée afin de mieux comprendre la distance de Fréchet est l'exemple cité par [Devogele, 1997] d'un maître et de son chien se déplaçant le long d'une ligne : "ils avancent et ils s'arrêtent indépendamment à volonté. La distance de Fréchet entre les deux lignes et la longueur minimale de la laisse qui permet la progression simultanée".
Plus la distance de Fréchet entre deux strokes est faible, plus ce critère est favorable à l'appariement.

0.2.2.1.3 Critère d'orientation Nous calculons l'écart d'orientation entre deux strokes en mesurant l'angle entre les directions globales des deux strokes. L'orientation générale d'un arc est définie à partir des contributions de l'orientation de chacun de ses segments à l'orientation globale [Hangouët, 1998] (c.f figure8). Plus un segment est long, plus la contribution de son orientation sera importante. Plus la différence d'orientation est faible,plus ce critère est favorable à l'appariement des deux candidats.

0.2.2.1.4 Critère de proximité globale L'utilisation seul d'un critère de distance ne suffit pas, car deux géométries globalement proches mais s'éloignant par exemple en une extrémité peuvent avoir une distance de Fréchet importante alors que les strokes se suivent plutôt bien dans l'ensemble. Ce critère perrmet de s'assurer que deux géométries ne s'éloignent pas trop l'une de l'autre globalement. Il se base sur l'indicateur développé par [Goodchild and Hunter, 1997] pour mesurer les incertitudes des écarts géométriques des polylignes en mesurant le taux d'inclusion d'une polyligne à contrôler dans la bande ϵ (buffer de rayon ϵ) de la polyligne de référence [Bel Hadj Ali, 2001].
Nous construisons un buffer autour des strokes candidats à l'appariement. Le rayon de ce buffer prend en compte les déformations locales des fonds de cartes et est fonction des écarts constatés lors de leur

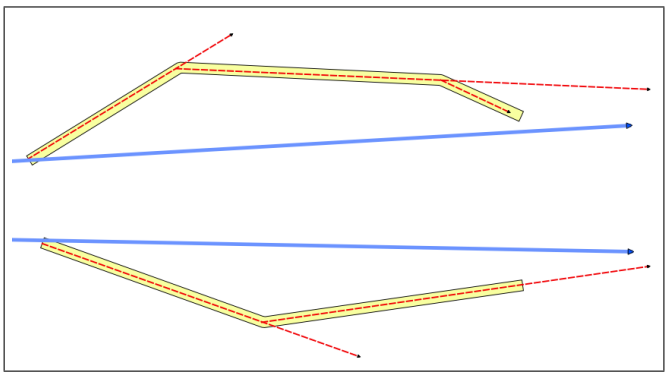


FIGURE 8: Orientation générale d'un arc.

géoréférencement. Soient $BUF1$ et $BUF2$ les buffers construits autour des strokes à apparier s_1 et s_2. Notons $|s|$ la longueur du stroke s. Soit $sint_1 = s_1 \cap BUF2$ la géométrie résultant de l'intersection de s_1 avec $BUF2$, et $sint_2 = s_2 \cap BUF1$ celle de l'intersection de s_2 avec $BUF1$. La valeur mesurée pour le critère de recouvrement vaut alors : $max(\frac{|s_1|}{sinter_1}, \frac{|s_2|}{sinter_2})$, où $\frac{s_i}{sinter_i}$ représente la proportion de la longueur du stroke s_i comprise dans le buffer d'intersection.

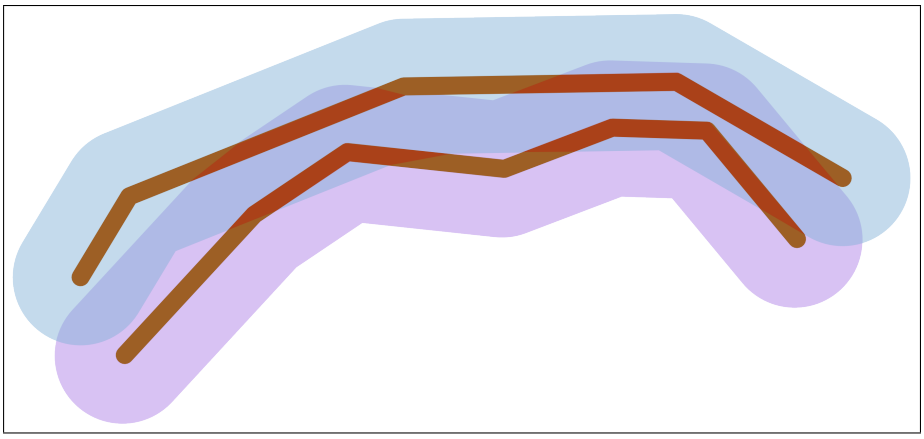


FIGURE 9: Calcul du critère de proximité globale : les zones plus sombres correspondent aux intersections des géométries linéaires avec les buffers.

0.2.2.1.5 Critère toponymique Peu de cours d'eau sont nommés sur les feuilles de Cassini. Seuls les toponymes des fleuves principaux sont renseignés, à l'exception de la feuille de Saint-Malo. Nous utilisons une mesure de distance toponymique qui permet d'aider à l'appariement de ces cours d'eau majeurs. La distance utilisée est basée sur la distance de Damerau-Levenshtein [Levenshtein, 1965, ; Damerau, 1964] et reprend les principes de la distance définie dans [Samal et al., 2004] qui permet le calcul d'un coefficient de similarité entre deux chaînes de caractères.

0.2.2.2 Prise de décision et détermination des liens d'appariement

Chaque stroke possède potentiellement plusieurs candidats à l'appariement. Nous présentons deux méthodes permettant de prendre une décision, c'est-à-dire de choisir quel candidat(s) retenir. L'une calcule une fonction de score et compare la valeur obtenue à un seuil de prise de décision, pouvant être définit automatiquement à l'aide d'une métaheuristique. La seconde repose sur l'utilisation d'une AHP (Analytic Hierarchy Process), une technique d'analyse multi-critères mise au point par Thomas L. Saaty dans les années 1970 [Saaty, 2008].

Avant toute opération de fusion de mesures et de prise de décision, il convient de normaliser chaque mesure dans l'intervalle $[0, 1]$ afin de pouvoir les comparer. Nous appliquons ainsi à chaque mesure de critère une fonction reflétant le degré avec lequel nous croyons en l'appariement d'un objet avec son candidat pour la valeur mesurée du critère [Olteanu, 2008 ; Costes et al., 2012]. Il est montré dans [Costes, 2012] que le choix d'un type de courbe (exponentielle, logarithmique, linéaire, etc.) influe sur la qualité de l'appariement. Il est suggéré de choisir le type de courbe qui reflète l'attitude du critère attendue par l'utilisateur. Par exemple, il est difficile de privilégier un candidat plutôt qu'un autre lorsque leur distance géométrique à l'objet étudié est inférieure à la précision du fond de carte, estimée à plusieurs centaines de mètres. L'utilisation d'une courbe exponentielle pour le critère géométrique permet de modéliser cette difficulté. Lorsque nous ne savons pas par quelle courbe modéliser un critère, l'utilisation d'une fonction linéaire est le choix le plus prudent.
Nous choisissons comme fonctions de normalisation :
- une fonction exponentielle pour la distance de Fréchet,
- une fonction exponentielle pour le critère d'orientation : vu la précision des données, des écarts d'orientation faible ne permettent pas de privilégier un candidat par rapport à l'autre. Par contre, au delà d'un seuil, la différence devient rédhibitoire,
- une fonction linéaire pour les trois autres critères, car leur comportement est délicat à anticiper.

0.2.2.2.1 Prise de décision par utilisation d'une fonction de score et comparaison à un seuil de prise de décision

Fonction de score Dans un contexte d'appariement, un critère peut avoir plus d'importance qu'un autre pour trier les candidats. Nous associons donc à chaque critère C_i un poids p_i qui représente sont degré d'importance global dans la prise de décision.
Nous définissons une fonction de score F, calculée pour chaque candidat à l'appariement comme une pondération des mesures des critères par les poids précédemment définis. Si c est un candidat à l'appariement et m_i la mesure du i^{me} critère, alors le score du candidat vaut : $\sum_{i=1}^{5} p_i * m_i$

Prise de décision Le score d'un candidat est ensuite comparé à un seuil permettant la prise de décision : en deçà d'une certaine valeur, l'algorithme déclare l'appariement non valide.
Un des problème posé par cette approche réside dans le choix du seuil de prise de décision, souvent faîte empiriquement, ou par apprentissage sur des données appariées manuellement. Nous proposons ici une méthode automatique de calcul du seuil de précision. Notons bien que l'algorithme ne garantit pas de fournir la solution optimale, mais bien une solution approchée de la solution optimale.
Nous assimilons la problématique de détermination automatique du seuil de prise de décision à un problème d'optimisation difficile, pour lequel nous cherchons une solution réalisable en un temps raisonnable. Explorer l'espace des solutions dans son ensemble n'est pas une méthode envisageable, pour des raisons évidentes d'explosion du temps de calcul : si les deux réseaux à apparier sont composés de n strokes, il y aurait potentiellement $(n!)^2$ couples candidats à analyser en supposant que chaque stroke est apparié une seule fois. Même en restreignant l'espace de recherche, ce procédé s'avère temporellement trop complexe.
Nous utilisons une métaheuristique, et plus précisément un algorithme génétique donnant une solution approximative sous forme de liens d'appariements en un temps polynomial. Nous prenons alors comme valeur de seuil de prise de décision pour l'algorithme d'appariement hiérarchique, la moyenne des scores pondérés pour la solution obtenue par l'algorithme génétique.
Nous détaillons brièvement l'approche proposée.
Les algorithmes génétiques appartiennent à la famille des algorithmes évolutionnaires et utilisent la notion de sélection naturelle. Ils permettent de trouver des solutions à des problèmes d'optimisation en un temps raisonnable. Pour notre problème, un individu est un lien d'appariement, et possède deux gènes, représentant l'identifiant du stroke de référence (Cassini) et du stroke candidat à l'appariement (BDCarto). Nous utilisons un codage binaire pour les gènes. Prenons l'exemple de la feuille de Grenoble.

Les réseaux hydrographiques de Cassini et de la BDCarto comprennent respectivement 224 et 132 strokes. Nous codons donc chaque gène de chaque individus sur 8 bits ($2^8 = 256 > 132$, et $2^7 = 128 < 132$). Par exemple, le chromosome suivant : $x = \{01111101, 00101011\}$ représente le lien d'appariement entre les strokes d'identifiant respectivement 125 et 43. Par la suite, nous notons C1 la contrainte : "le code binaire d'un gène doit correspondre à l'identifiant d'un stroke existant dans la base" et C2 la contrainte : "le stroke de référence et le candidat à l'appariement doivent être suffisamment proches", qui permet de filtrer les candidats parasites trop éloignés.
L'algorithme génétique implémenté comprend 5 étapes :

- Étape 1 : création de la population initiale. Les individus sont crées aléatoirement en respectant les contraintes C1 et C2. La taille de la population initiale est choisie empiriquement, en supposant qu'on aura en sortie au maximum autant de liens d'appariement qu'il y a de strokes dans la base de données de référence,
- Étape 2 : évaluation de la population. La fonction de *fitness* (qu'on cherche à maximiser) utilisée ici est la fonction de score détaillé dans la section précédente,
- Étape 3 : Sélection d'une partie de la population susceptible d'évoluer, par tournoi stochastique binaire sans remise du vainqueur. On choisit au hasard deux individus qui sont évalués (ils "combattent)" : le meilleur (celui dont le score, i.e la fonction de *fitness*, est le plus élevé) l'emporte avec une probabilité comprise entre 0.5 et 1. Le perdant est replacé dans la population, et on réitère le processus jusqu'à avoir sélectionné suffisamment d'individus. La sélection par tournoi est intéressante car elle limite les effets de convergence prématurée,
- Étape 4 : croisement et mutation. Cette étape consiste à "mélanger" et faire muter le bagage génétique des individus sélectionnés afin de faire apparaître de nouveaux individus qui seront potentiellement meilleurs. Le croisement est réalisé par enjambement en un point : un couple d'individus sélectionnés est choisi au hasard et une position (point de recombinaison) dans la chaîne génétique (la concaténation du code binaire des deux gênes) est tirée aléatoirement. Deux individus fils sont alors créés par interversion des codes génétiques des parents par rapport au point de recombinaison. Considérons par exemple les individus $x_1 = \{00001111\}$ et $x_2 = \{11110000\}$. Supposons que le point de recombinaison choisi au hasard correspondent à la position 3 dans la chaîne. Les individus fils engendrés par le croisement seront alors : $x_3 = \{00110000\}$ et $x_4 = \{11001111\}$. La mutation consiste à changer un bit du code génétique d'un individu sélectionné avec une probabilité généralement comprise entre 0.01% et 1%. La position du bit modifié est choisie aléatoirement. Appliquons par exemple une mutation à l'individu x_2 en position 2. L'individu muté sera : $x_5 = \{10110000\}$,
- Étape 5 : insertion des nouveaux individus dans la population. Nous insérons les nouveaux individus dans la population originelle, à l'exception de ceux ne respectant pas les contraintes C1 et C2. Les individus enfants sont donc mélangés avec leurs parents. Nous décidons de garder constant le nombre d'individus de la population à chaque itération, et supprimons donc les individus en trop ayant le score le plus faible. Puis le processus reprend à l'étape 2.

Nous stoppons l'algorithme lorsque la population n'évolue plus, *i.e* lorsque le score moyen de ses individus est stable depuis une certain nombre de générations (*e.g* 20 itérations). Ce score moyen est utilisé comme seuil de prise de décision dans l'algorithme d'appariement hiérarchique.

0.2.2.2.2 Prise de décision utilisant une AHP

L'AHP est un outil performant dans le cadre de la prise de décisions complexe ([Saaty, 2008]), basé sur une analyse hiérarchique de différents critères et candidats permettant d'atteindre un objectif (prendre une décision). Nous ne détaillons pas ici le principe théorique de fonctionnement de l'AHP, mais l'illustrons dans notre cadre d'appariement. L'approche s'effectue en 4 temps.

Regroupement des candidats Tous les candidats à l'appariement d'un objet étudié sont sélectionnés afin d'être comparés parallèlement. Nous construisons également une alternative fictive prise en compte dans l'analyse : "l'objet n'est pas apparié", qui devra être choisie si aucun candidat ne se démarque particulièrement en sortie de l'AHP. Nous notons "NAP" cette alternative.

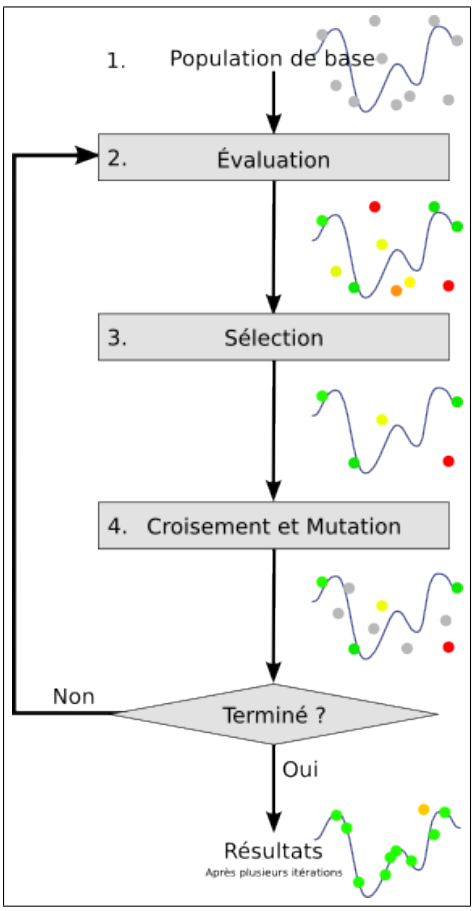


FIGURE 10: Schéma de fonctionnement d'un algorithme génétique (source Wikipédia)

Confrontation des candidats Pour chaque critère, les scores des candidats sont confrontés deux à deux. Le résultat de la comparaison est calculé selon le principe de "the Fundamental Scale for Pairwise Comparisons" de l'AHP, qui attribut une importance sous la forme d'un nombre aux deux candidats comparés, fonctions de leur score, et qui dépend de l'ordre dans lequel les candidats sont étudiés : si $c1$ et $c2$ sont deux candidats, et si nous notons f la fonction de comparaison, alors $f(c_1, c_2) = \frac{1}{f(c_2,c_1)}$. Lors de la confrontation d'un candidat ayant comme score m_i à l'hypothèse "NAP", nous associons à celle-ci le score $1 - m_i$, rendant ainsi compte dans quelle mesure nous ne croyons pas que ce candidat soit apparié pour sa valeur de score mesurée. Le résultat de cette comparaison est entré dans une matrice de taille $n * n$ où n est le nombre de candidats confrontés. Un vecteur de priorités est ensuite calculé rendant compte de l'importance relative des candidats pour ce critère, dérivés du jugement établit par la comparaison précédente. Mathématiquement, ce vecteur correspond au vecteur propre principal de la matrice, *i.e* le vecteur propre associé à la plus grande valeur propre en valeur absolue. A chaque matrice de comparaison est associé un nombre, le ratio de consistance (CR) de la matrice, qui donne une indication sur l'inconsistance faîte dans le jugements des confrontations. En pratique, on cherche à s'assurer que $CR \leq 0.10$ afin de garantir la bonne cohérence de la comparaison ([Saaty, 2008]).
Nous donnons en exemple de construction de cette matrice pour deux candidats à l'appariement, pour un critère arbitraire donné.

TABLE 2: Exemple de construction de la matrice de comparaison et des priorités pour deux candidats et l'alternative "NAP", pour un critère donné.

Dans notre exemple, c'est le candidat 1 qui l'emporte largement pour le critère étudié.

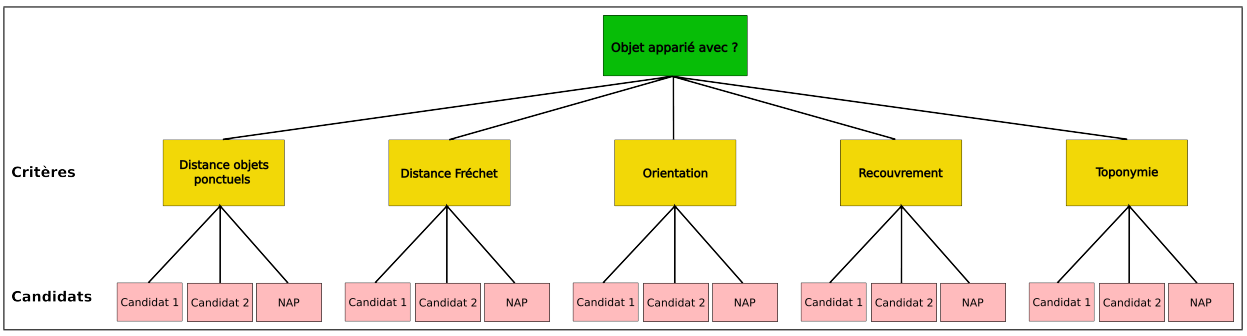


FIGURE 11: Un exemple de hiérarchie d'une AHP dans le cadre de l'appariement de données : le but à atteindre est de choisir un candidat à l'appariement avec un objet donné. Dans notre exemple, il y a deux candidats potentiels, plus l'hypothèse "l'objet n'est pas apparié"

Intensity of Importance	*Definition*	*Explanation*
1	Equal Importance	Two activities contribute equally to the objective
2	Weak or slight	
3	Moderate importance	Experience and judgement slightly favour one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgement strongly favour one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	An activity is favoured very strongly over another; its dominance demonstrated in practice
8	Very, very strong	
9	Extreme importance	The evidence favouring one activity over another is of the highest possible order of affirmation
Reciprocals of above	If activity *i* has one of the above non-zero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i*	A reasonable assumption
1.1–1.9	If the activities are very close	May be difficult to assign the best value but when compared with other contrasting activities the size of the small numbers would not be too noticeable, yet they can still indicate the relative importance of the activities.

FIGURE 12: "The Fundamental Scale of Pairwise Comparisons" ([Saaty, 2008]) permet de comparer deux candidats à l'appariement et d'attribuer un degré d'importance d'un candidat par rapport à l'autre, pour les critère considéré.

Candidats	**Score**	**Score NAP**
candidat 1	0.65	0.35
candidat 2	0.45	0.55

Matrice de comparaison				**Priorités**
	Candidat 1	Candidat 2	NAP	
Candidat 1	1	3	4	0.63
Candidat 2	$\frac{1}{3}$	1	$\frac{1}{2}$	0.15
NAP	$\frac{1}{4}$	2	1	0.22
				CR = 0.09

Confrontation mutuelle des critères Contrairement à la méthode précédente qui attribuait un poids global à un critère dans la prise de décision, l'AHP permet de pondérer chaque critère en fonction de l'importance que nous lui accordons par rapport à chaque autre critère. Autrement dit, si dans l'approche par fonction de score, nous disposions d'un cinq-uplet de poids, l'AHP nous permet de définir une matrice de poids de taille $5 * 5$ ce qui ajoute beaucoup de finesse dans la définition des importances relatives des critères.
À partir de la matrice de confrontation des critères, un vecteur de priorités et un ratio d'inconsistance sont calculés comme précédemment.

Prise de décision L'idée est ici d'obtenir pour chaque candidat une priorité globale, combinant les priorités du candidat pour chaque critère et tenant compte des priorités relatives des critères les uns par rapport aux autres. Soient $(C_1, ..., C_N)$ N candidats. Notons $(\alpha_{i,1}, ..., \alpha_{i,N})$ le vecteur de priorités

Critères	Distance obj. ponctuels	Distance Fréchet	Orientation	Recouvrement	Toponymie	**Priorités**
Distance obj. ponctuels	1	2	1	1	$\frac{1}{3}$	0.17
Distance Fréchet	$\frac{1}{2}$	1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{3}$	0.07
Orientation	1	4	1	$\frac{1}{2}$	1	0.21
Recouvrement	1	4	2	1	1	0.27
Toponymie	3	3	1	1	1	0.28
						CR = 0.05

TABLE 3: Exemple de matrice de confrontation des critères. On constate la prédominance des critères de toponymie et de recouvrement.

des candidats pour le i^{me} critère. Enfin, soit $(p_1, ..., p_5)$ le vecteur de priorités issus de la confrontation des critères. La priorité globale $P(C_j)$ du candidat C_j est donnée par :

$$P(C_j) = \sum_{i=1}^{5} p_i * \alpha_{i,j}$$

Nous choisissons comme solution le candidat ayant la plus grande priorité globale. Si c'est l'alternative "NAP" qui est ici choisie, alors l'objet est non apparié.

	Distance obj. ponctuels	Distance Fréchet	Orientation	Recouvrement	Toponymie	**Priorité globale**
Candidat 1	0.15	0.90	0.1	0.03	0.15	**0.1596**
Candidat 2	0.15	0.05	0.85	0.02	0.05	**0.2269**
NAP	0.70	0.05	0.05	0.95	0.80	**0.6135**

TABLE 4: Exemple de prise de décision pour deux candidats. Les colonnes contiennent les priorités des candidats pour chaque critère. Les priorités issues de la confrontation des critères entre eux sont celle du tableau 3. Malgré une faible distance entre l'objet à apparier et le candidat 1, et malgré une différence d'orientation faible entre l'objet et le candidat 2, c'est l'hypothèse NAP qui est retenue. L'objet n'est donc pas apparié.

Nous donnons le détail du calcul de la priorité du premier candidat de l'exemple du tableau 5 en guise d'illustration : $P(candidat_1) = 0.15 * 0.17 + 0.90 * 0.07 + 0.1 * 0.21 + 0.03 * 0.27 + 0.15 * 0.28 = 0.1596$.

0.2.2.3 Traitement des strokes non appariés Les strokes non appariés sont ensuite traités séparément, et pour chacun d'entre eux nous appliquons l'algorithme d'appariement de [Mustière and Devogele, 2008]. Afin de gagner en temps de calcul, nous nous restreignons au sous-réseau constitué du stroke étudié, mais également des strokes qui lui sont connectés afin d'éviter d'introduire des effets de bord vu la la simplification topologique qui est effectuée ici. Les liens issus de cette étape sont qualifiés d'incertains et nécessitent une vérification manuelle en post-traitement.

0.2.3 Évaluation des résultats

Il est délicat d'évaluer la qualité de l'algorithme d'appariement. En effet, les réseaux présentent de grandes différences d'organisation comme nous l'avons signalé précédemment. Beaucoup de strokes sont impossibles à apparier manuellement. Il est donc compliqué, dans certains cas, de juger de la justesse d'un lien établi ou manquant. On remarque également que plus l'ordre d'un stroke est élevé, plus l'identification des correspondants est complexe. Pour donner une évaluation quantitative de l'appariement, nous avons réalisé un appariement manuel sur les strokes pouvant être appariés manuellement sans ambiguïté.
Les liens d'appariement en sortie sont visuellement bons. Les grandes structures (bras principaux des fleuves, rivières, ruisseaux importants, etc.) sont correctement mis en correspondance. Le tableau ci-

dessous détaille les scores obtenus sur les trois zones en utilisant les deux méthodes de prise de décision.

	liens d'appariement			objets non appariés		
Zones	**Précision**	**Rappel**	**FScore**	**Précision**	**Rappel**	**FScore**
Prise de décision par fonction de score						
Reims	95%	86%	90%	45%	80%	57%
Grenoble	82%	80%	81%	94%	91%	92%
St-Malo	77%	86%	81%	100%	94%	97%
Prise de décision par AHP						
Reims	86%	95%	90%	67%	80%	73%
Grenoble	80%	92%	86%	95%	90%	92%
St-Malo	68%	93%	78%	100%	88%	94%

TABLE 5: Récapitulatif des résultats quantitatifs de l'évaluation de l'approche d'appariement hiérarchique multi-critère dont la prise de décision est faîte dans un premier temps par comparaison d'une fonction de score à un seuil préalablement défini, puis basée sur une AHP

Notre approche permet d'améliorer sensiblement les résultats obtenus par les méthodes basées uniquement sur ces critères géométriques et topologiques et les approches par analyse multi-critères non hiérarchiques (c.f 1).
Les scores des liens d'appariement sont moins bons sur Saint-Malo, car il s'agit de la zone sur laquelle le réseau hydrographique a le plus changé. En effet, aucuns des grands fleuves cartographiés à l'époque de Cassini sur les zones sableuses n'a d'homologue actuel. Ces changements majeurs dans la structure du réseau se répercutent par la détection de nombreux faux positifs par tous les processus d'appariement testés.
Le FScore des objets non appariés sur la zone de Reims est plus faible que sur les autres zones. Il s'agit en fait d'une sensibilité induite par le très faible nombre d'objets n'ayant pas d'homologue actuel (5 entités). La plupart de ces objets sont bien non appariés par le processus automatique (4 objets sur 5 d'où un rappel de 80%), mais celui-ci n'apparie pas certains cours d'eau qui auraient du l'être (faux négatifs). Même si le nombre de faux négatifs est faible, le nombre de vrais négatifs étant également faible implique une précision moyenne pour la détection des objets non appariés.
L'appariement hiérarchique par prise de décision utilisant une AHP donne des résultats similaires en moyenne à l'appariement hiérarchique par prise de décision par comparaison d'une fonction de score à un seuil pour les liens d'appariement (86%), mais est légèrement meilleur en moyenne concernant les objets non appariés (86% contre 82%).

0.3 Conclusion, limites et perspectives

Les approches d'appariements de géométries linéaires issues de la littérature ne permettent pas de gérer les réseaux présentant à la fois d'importants décalages géométriques et topologiques, mais également des imperfections majeures : imprécisions du tracé, incertitude et incomplétude de la toponymie et de la sémantique, etc. Nous avons proposé une nouvelle approche d'appariement adaptée aux données anciennes et testée sur deux bases de données présentant une différence de temporalité de près de 250 ans : le réseau hydrographique issus des fonds de cartes de Cassini, et la BDCarto, base de données topographique de référence actuelle. L'approche présentée est hiérarchique et itérative : l'organisation du réseau est reconstruite en reconstituant les structures de bonne continuité naturelle (strokes), représentant réellement les entités hydrographiques telles que les rivières et les fleuves. L'agencement hiérarchique ainsi défini sert de pivot à l'appariement, en ne retenant parmi les candidats potentiels que ceux dont les parents, ou eux-même, ont été préalablement appariés. L'approche est également

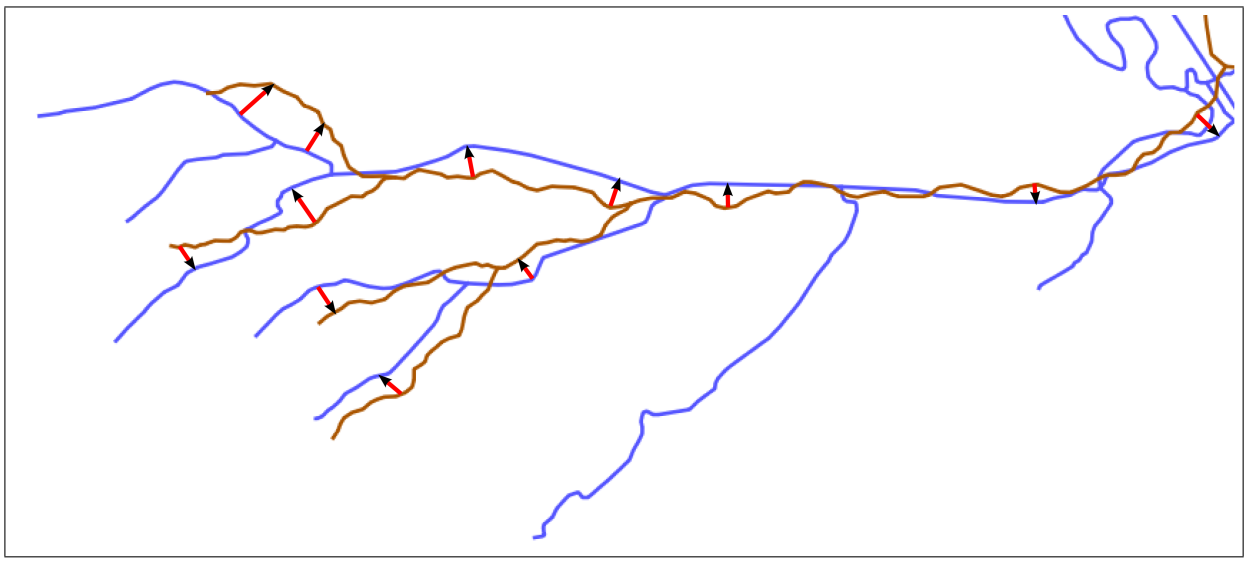


FIGURE 13: Résultat d'appariement hiérarchique sur la zone de Reims, sans aucun faux positifs ni faux négatifs.

multi-critères, et juge les candidats à l'appariement en fonction de leur distance à l'objet, de leur orientation générale, de leur proximité globale, de leur nom. Nous proposons de plus l'introduction d'un nouveau critère, utilisant un appariement effectué en amont d'entités ponctuelles vectorisées des fonds de cartes de Cassini (églises, hameaux, moulins, etc.), et basé sur l'hypothèse que le lits des cours d'eau a peu changé depuis le 18^{me} siècle, en mesurant la proportion d'objets ponctuels précédemment mis en correspondance et proches des cours d'eau candidats à l'appariement. Enfin, nous proposons deux méthodologies permettant la prise de décision. La première est basée sur le calcul d'une fonction de score pondérant chaque mesure de critère et sa comparaison à une valeur seuil. Nous présentons un algorithme de détermination automatique d'un seuil approximatif de prise de décision, fondé sur un algorithme génétique. La seconde méthode repose sur l'utilisation d'une technique de décision multi-critère très utilisée dans l'industrie : l'AHP. Les résultats obtenus par nos deux approches permettent d'améliorer sensiblement la qualité de l'appariement des réseaux.
Cependant, nous pensons que notre approche hiérarchique est principalement adaptée aux réseaux "presque" en arbre, c'est-à-dire aux réseaux pour lesquels il existe quasiment tout le temps un unique chemin pour relier deux points. En effet, réaliser une classification de [Horton, 1945] sur un réseau type réseau routier n'a pas de sens, car le résultat de la classification dépendrait de l'ordre dans lequel les strokes sont analysés. Afin d'étendre l'appariement multi-critère hiérarchique à d'autres types de réseaux, il convient de définir autrement l'ordre des strokes (indices de centralité, longueur du stroke, etc.).

0.4 L'appariement du réseau routier

Nous avons utilisé l'approche de [Mustière and Devogele, 2008] pour apparier les réseaux routiers. Les résultats sont mitigés, étant donné les grandes variations de niveaux de détail, de précision, de géométrie et de topologie entre les deux réseaux. Quelques routes ne sont pas appariables car aucun candidat ne se distingue particulièrement. Cependant, le tracé de certaines routes principales suit particulièrement bien la géométrie des départementales actuelles.

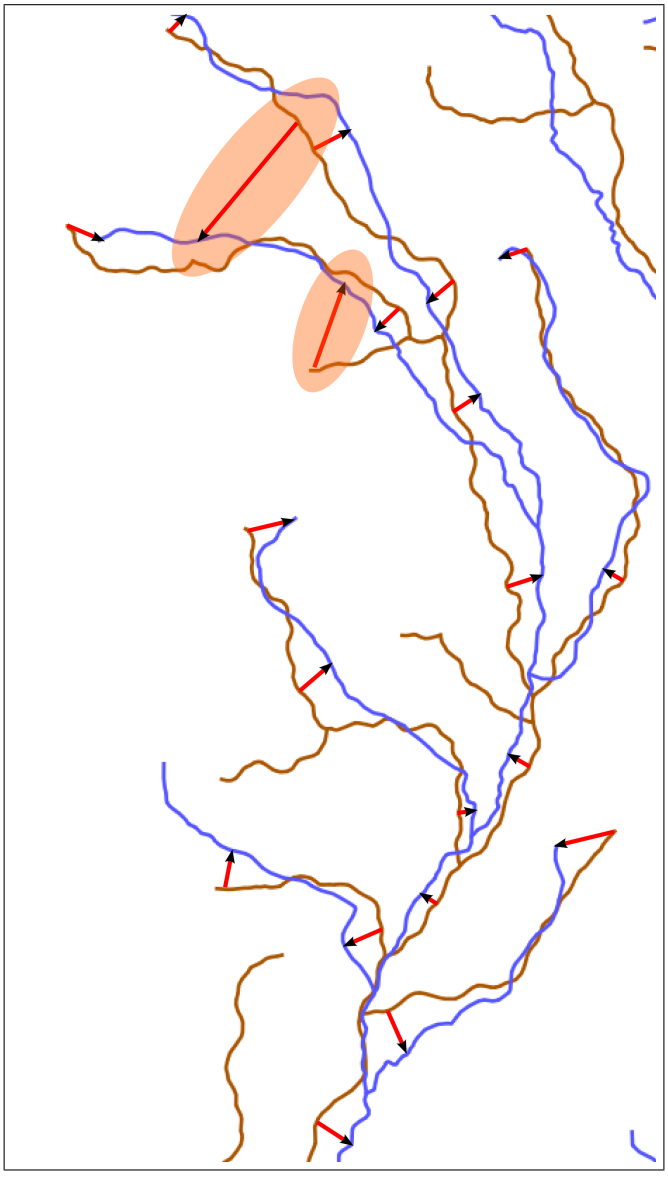


FIGURE 14: Résultat d'appariement hiérarchique sur la zone de Grenoble. Deux faux positifs sont détectés (tâches orange).

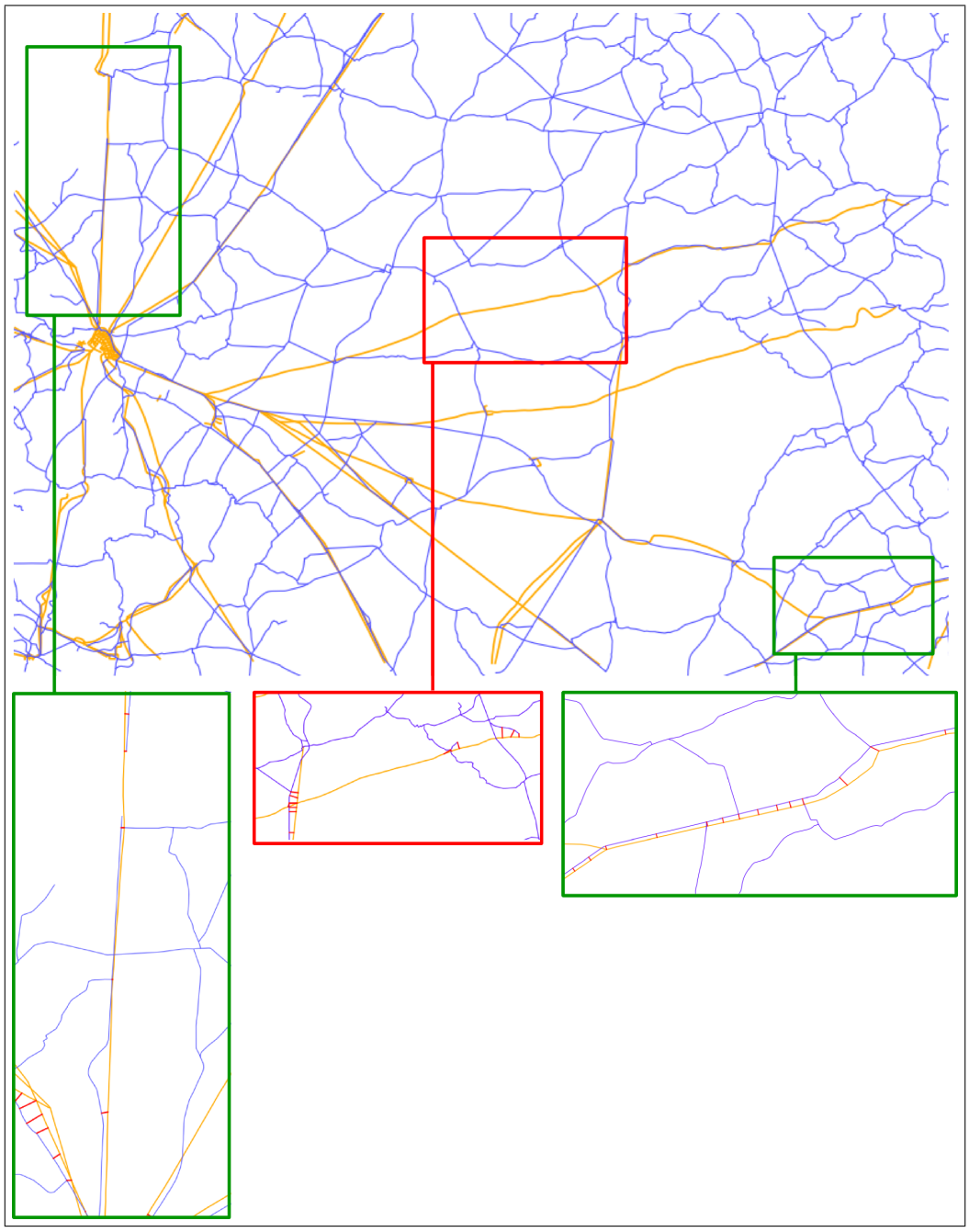


FIGURE 15: Exemples d'appariement réussis (cadres verts) et d'erreur d'appariement (cadre rouge) pour le réseau routier de la zone de Reims.

Bibliographie

Atef Bel Hadj Ali. *Qualité géométrique des entités géographiques surfaciques. Application à l'appariement et définition d'une typologie des écarts géométriques.* PhD thesis, Université de Marne-La-Vallée, 2001.

B. Costes. Appariement cassini - rge (géométries ponctuelles) v1 (l2.3-1). Rapport technique, Projet ANR GéoPeuple, Institut Géographique National, Laboratoire COGIT, 2012.

B. Costes, E. Grosso, and C. Plumejeaud. Géoréférencement et appariement de données issues des cartes de cassini. In *SAGEO*, 2012.

Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3) :171–176, March 1964.

Thomas Devogele. *Processus d'intégration et d'appariement de bases de données Géographiques, Applications á une base de données routières multi-échelles.* PhD thesis, Université de Versailles, 1997.

Heikk * Tech. Report CD-TR 94/64 Christian Doppler Laboratory for Expert Systems * TU Vienna Austria Eiter, Thomas andMannila. Computing discrete fréchet distance. Technical Report Tech. Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, 1994.

M.F Goodchild and G Hunter. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Systems*, 11(3) :299–306, 1997.

J.F Hangouët. *Approche et méthodes pour l'automatisation de la généralisation cartographique ; application en bord de ville.* PhD thesis, Université de Marne-La-Vallée, 1998.

R.A. Horton. Erosional development of streams and their drainage basins : hydrophisical approach to quantitative morphology. *Geological Society of America Bulletin*, 56 :275–370, 1945.

Bin Jiang, Sijian Zhao, and Junjun Yin. Self-organized natural roads for predicting traffic flow : A sensitivity study. *Journal of Statistical Mechanics : Theory and Experiment*, July 2008.

V Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 4(163) :845–848, 1965,.

P. Lüscher, D. Burghardt, and R. Weibel. Matching road data of scales with an order of magnitude difference. In *23th International Cartographic Conference*, 2007.

S. Mustière and T. Devogele. Matching networks with different levels of detail. *GeoInformatica*, 12 : 435–453, 2008.

A.M. Olteanu. *Fusion de connaissances imparfaites pour l'appariement de donnï¿œes gï¿œographiques. Proposition d'une approche s'appuyant sur la thï¿œorie des fonctions de croyance.* PhD thesis, Université Paris-Est, 2008.

Thomas L Saaty. Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, 1(1), 2008.

A Samal, S Seth, and K Cueto. A feature-based approach to conflation of geospatial sources. In *International Journal of Geographical Information Sciences*, volume 18, pages 459–489, 2004.

R. Thomson and R. Brooks. Efficient generalization and abstraction of network data using perceptual grouping. In *5th GeoComputation*, 2000.

R. Thomson and D. Richardson. The 'good continuation' principle of perceptual organization applied to the generalization of road networks. In *19th International Cartographic Conference*, 1999.

G. Touya. Enrichissement automatique de donées par analyse spatiale pour la généralisation de réseaux. *RIG*, 20, 2010.

V. Walter and D. Fritsch. Matching spatial data sets : a statistical approach. *International Journal of Geographical Information Science*, 13 :5 :445–473, 1999.

M Zhang, W Shi, and L.A Meng. A generic matching algorithm for line networks of different resolutions. In *ICA Workshop on Generalisation and Multiple Representations*, 2005.

0.5 Annexe

Algorithm 1 Sélectionne le segment suivant à partir d'un segment donné, d'après notre méthode

```
1: currentSegment le segment en cours de traitement
2: thresold l'angle de déflexion maximal
3: direction la directin (from, ou to)
4: procedure constructStroke(currentSegment, direction)
5:     if direction is 'from' then
6:         searchPoint = the from point of currentSegment ;
7:     else if direction is 'to' then
8:         searchPoint = the to point of currentSegment ;
9:     end if
10:    search the segments intersected with searchPoint except currentSegment ;
11:    if there are no intersected segments then
12:        return currentSegment ;
13:    end if
14:    if the searched segments are all already processed then
15:        return currentSegment ;
16:    end if
17:    Exclude the processed segments from the searched segments to get a remained set ;
18:    if currentSegment has toponym then
19:        return checkAttributeContinuity(currentSegment, "toponym") ;
20:    end if                                                    ▷ the segment has no toponym
21:    Exclude the segments which have the same toponym
22:    if there are no remaining segments then
23:        return currentSegment ;
24:    else if there is only one remaining segment then
25:        return that segment ;
26:    end if
27:    if currentSegment has nature attribute then
28:        return checkAttributeContinuity(currentSegment, "nature attribute") ;
29:    end if                                          ▷ the segment has no nature attribute
30:    Exclude the segments which have the same nature attribute
31:    if there are no remaining segments then
32:        return currentSegment ;
33:    else if there is only one remaining segment then
34:        return that segment ;
35:    end if
36:    Calculate the delfection angles (a1) bewteen currentSegment and every segments in the remained set ;
37:    Calculate the deflection angle (a2) of every pair in the remained set ;
38:    Select the segments wich meet with hte condition a1 < a2 ;
39:    if there are no selected segments then
40:        return currentSegment ;
41:    end if
42:    Get the minimum deflection mina1 angle from a1 and its corresponding segment ;
43:    if mina1 < thresold then
44:        Change the status of that segment to be processed ;
45:        return that segment ;
46:    else
47:        return currentSegment ;
48:    end if
49: end procedure
50: procedure checkAttributeContinuity(currentSegment, attribute)
51:    Select the segments with the same attribute to get a remained set ;
52:    if there are no remaining segments then
53:        return currentSegment ;
54:    else if there is only one remaining segment then
55:        return that segment ;
56:    else
57:        Calculate the delfection angles (a1) bewteen currentSegment and every segments in the remained set ;
58:        Calculate the deflection angle (a2) of every pair in the remained set ;
59:        Select the segments wich meet with hte condition a1 < a2 ;
60:        if there are no selected segments then
61:            return currentSegment ;
62:        end if
63:        Get the minimum deflection mina1 angle from a1 and its corresponding segment ;
64:        return that segment ;
65:    end if
66: end procedure
```