Mise en correspondance des terminaux

Rappel du travail :

Le travail demandé est de faire une jointure entre les deux fichiers suivants :

- « Terminals.csv » : fichier contenant la liste des terminaux extraites à partir des LLyods.
- « SMDG.csv » : fichier contenant la liste terminaux de SMDG.

Aucune clef primaire (identifiant) n'étant commun aux deux bases de données, nous avons donc décider de nous servir des coordonnées des terminaux présentes dans les deux fichiers pour effectuer un rapprochement entre les deux bases ; l'objectif étant de trouver pour chaque terminal du fichier « SMDG.csv » un terminal assez proche dans le fichier « Terminals.csv ».

Réalisation de la jointure entre les deux bases

Nous avons écrit un programme en python pour réaliser cette jointure. Nous avons décidé de garder les terminaux ayant une distance inférieur à 1000 mètres. Avec avoir exécuter notre programme et donc réalisé cette jointure, nous avons constaté que certains terminaux du fichier « SMDG.csv » étaient appariés plusieurs fois à des terminaux du fichier « Terminals.csv ».

Après analyse de quelques terminaux concernés, il semblerait que cela vient de l'absence du terminal dans le fichier « Terminals.csv » donc notre programme prend des terminaux un peu plus éloignés. Certains terminaux n'existent pas dans le fichier « Terminals.csv » ou ont changé de nom depuis l'extraction de cette base.

Nous avons ci-après deux exemples :

- L'un avec le port de Yokohama (Cf. Figure 1),
- L'autre avec le port de Cork (Cf. Figure 2).

Dans le premier exemple (celui du port de Yokohama), quatre terminaux du fichier « SMDG.csv » ont le même terminal le plus proche dans le fichier « Terminals.csv ». Cependant, il ne semble pas non plus y avoir d'autres possibilités : les autres terminaux du fichier « Terminals.csv » ne pouvant clairement pas être appareillé avec les terminaux du fichier « SMDG.csv ».

Dans le second exemple (celui du port de Cork), on voit qu'un terminal du fichier « Terminals.csv » est très proche de plusieurs terminaux du fichier « SMDG.csv ».

Au final, nous avons donc des terminaux du SMDG qui sont appariés plusieurs fois :

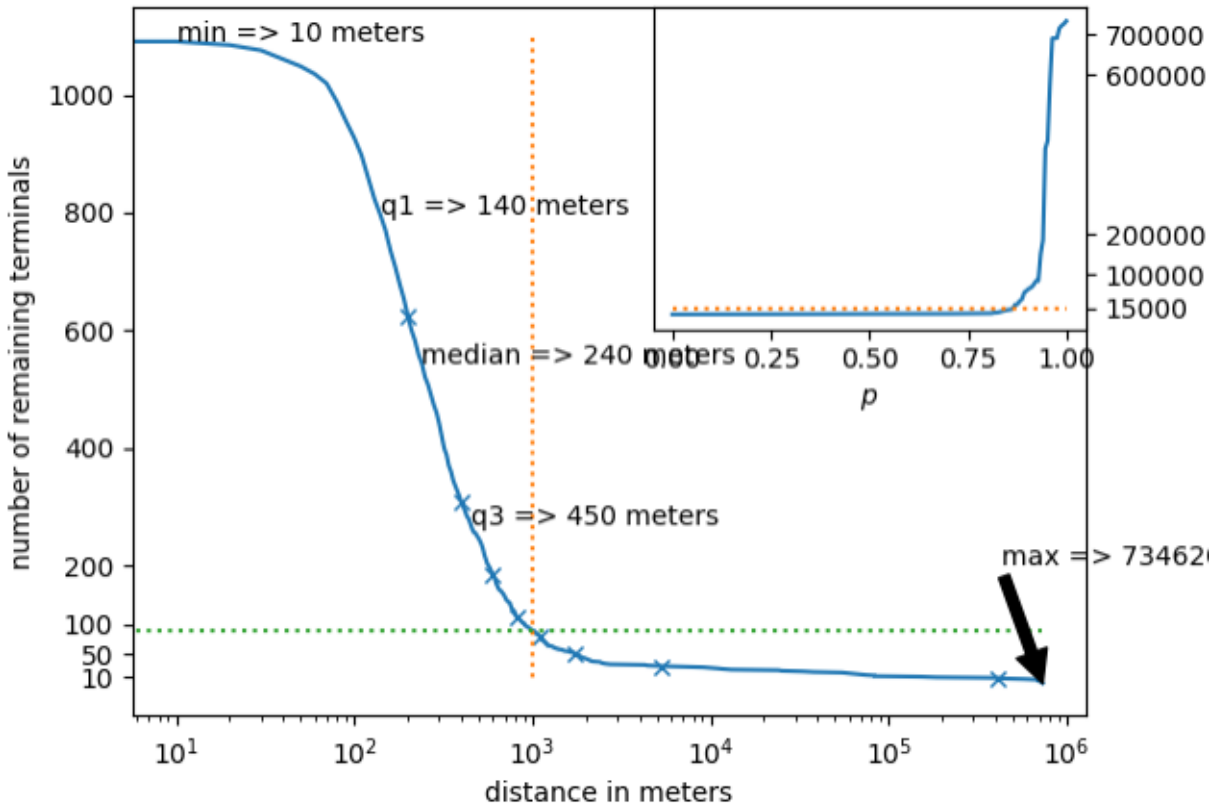
- 38 sont appariés deux fois ;
- 9 sont appariés trois fois ;
- Et 3 sont appariés quatre fois.

Figure 1 – Port de Yokohoma

Figure 2 - Port de Cork

Distance lors des appariements

Nous avons donc pris un seuil de 1000 mètres pour les appariement ce qui explique que certains terminaux (très peu) ne sont pas du tout appariés. Nous pouvons maintenant regarder à quelle distance se trouve majoritairement les terminaux appariés entre les deux bases grâce à la fonction de cumulative ci-dessous :



Sur le graphique ci-dessus, nous pouvons voir que 90 terminaux sont à une distance supérieur à 1000 mètres, soit environ 8 %. C'est pour cela que nous avons considérer que deux terminaux appareillés ne peuvent se trouver à une distance supérieure à 1000 mètres compte tenu de la taille des ports.

Résultat final

Pour la jointure des deux fichier nous avons un fichier CSV reprenant les données du fichier « SMDG.csv ». Nous avons ensuite ajouté une colonne nommée « tid » qui est l'identifiant du port dans le fichier « Terminals.csv ». Nous avons également ajouté les coordonnées du plus proches du terminal (dans le fichier Terminal.csv) dans la colonne « closest » et la distance calculé entre les coordonnées des terminaux dans les deux fichiers, dans la colonne « distance ». Pour repérer les terminaux avec incertitudes nous avons rajoutés la colonne « duplicate » qui permet d'indiquer combien de fois le plus proche terminal a été apparié.