

# Projet Data Visualisation

Chiara CORDIER  
Nolwenn DAVID  
Philippine RENAUDIN

# Théorie UMAP : résumé

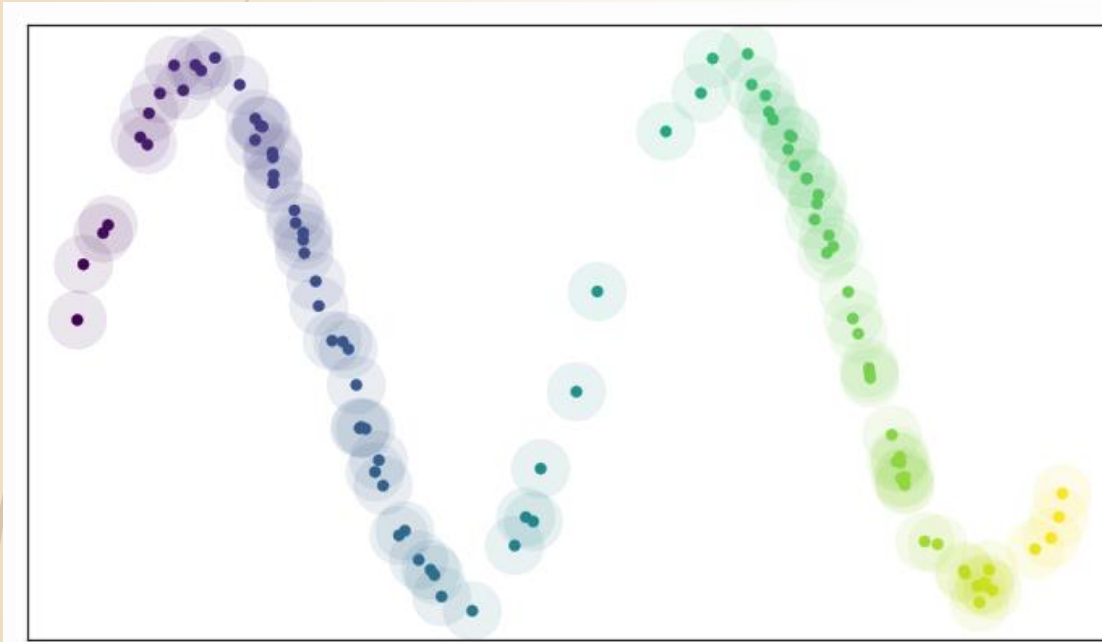
- Créateurs : Leland McInnes, John Healy, James Meville
- UMAP : Uniform Manifold Approximation and Projection
  - **réduction de dimension** linéaire et non linéaire (visualisation)
  - **Neighbour Graphs Techniques** (t-SNE)/Matrix factorization, pas de signification des axes
  - 3 hypothèses faites sur les données :
    1. Les données sont **distribuées de manière uniforme** sur une variété Riemannienne
    2. La métrique de Riemann est localement constante (ou peut être approximée comme tel)
    3. Les « points » sont localement connectés dans l'espace, il n'y a **pas de « points » isolés**
- Mathématiques derrière la méthode : algèbre topologique, Géométrie de Riemann, « logique floue »
- Représentation (simple) des données : trouvée en cherchant une projection sur un espace de faible dimension qui a une structure la plus proche possible de la structure topologique de départ.

# Au niveau mathématique

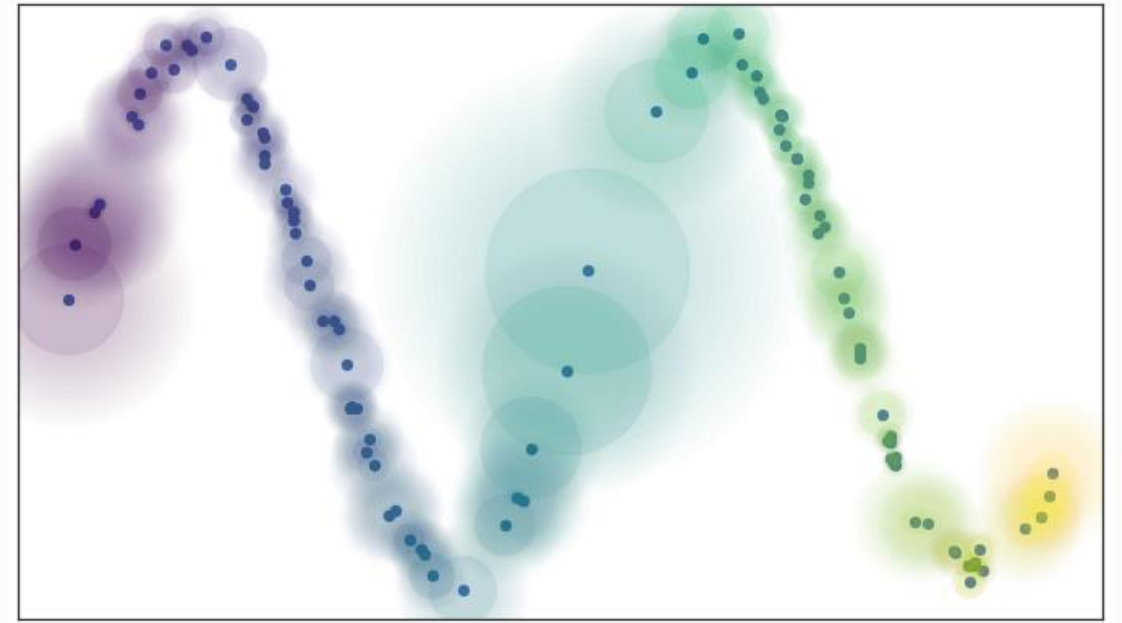
- ▶ Analyse de données topologiques : Il est possible en construisant d'une certaine manière des complexes simpliciaux dans un espace topologique de les **reconstruire de manière combinatoire sans perdre d'information** (on réussit à recouvrir toute l'information importante sur la topologie de l'espace de départ). Ce qui est plus simple à manipuler (typiquement  $\mathbb{R}^2$ )
- ▶ Hypothèse de **distribution uniforme** : Si les données ne sont pas uniformément distribuées sur la variété, on peut **définir une métrique** Riemannienne pour faire en sorte que l'hypothèse soit vérifiée, en faisant varier la notion de distance pour chaque type de données.

# Illustration

Même métrique pour tous les points. Toutes les sphères sont des sphères unités.



Métriques différentes. Les sphères sont également toutes des sphères unités mais en considérant chacune sa propre métrique.



Video explicative (anglais) : <https://www.youtube.com/watch?v=nq6iPZVUxZU>

# Au niveau algorithmique

- Construction d'une représentation topologique floue
  - Utilisation de l'entropie croisée

$$\sum_{e \in E} w_h(e) \log \left( \frac{w_h(e)}{w_l(e)} \right) + (1 - w_h(e)) \log \left( \frac{1 - w_h(e)}{1 - w_l(e)} \right)$$

- RP-trees (Rare Pattern Tree Mining)
  - NN-descent (Nearest Neighbor descent)
- Représentation dans un espace de faible dimension
  - SGD (Stochastique Gradient Descent)
  - Negative sampling

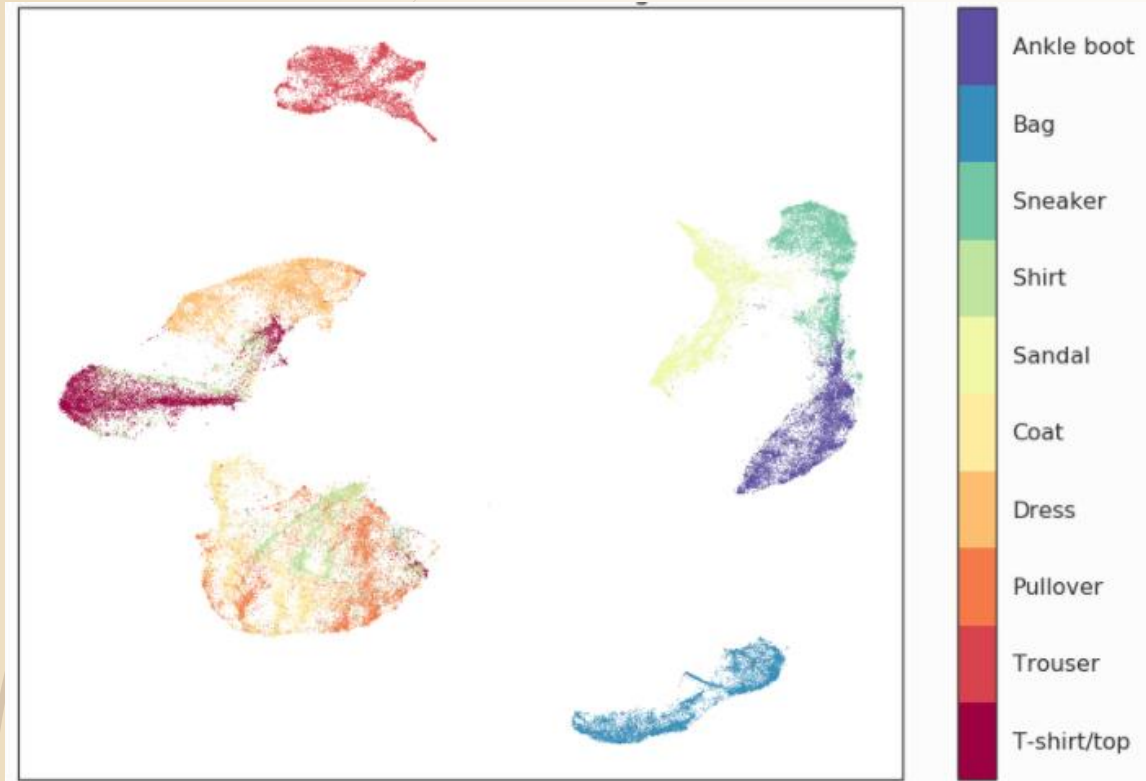


# Avantages

- Rapidité
- Bonne capture de la structure globale
- Peut faire de la classification supervisée comme non supervisée
- Peut prendre en charge plusieurs types de données en même temps

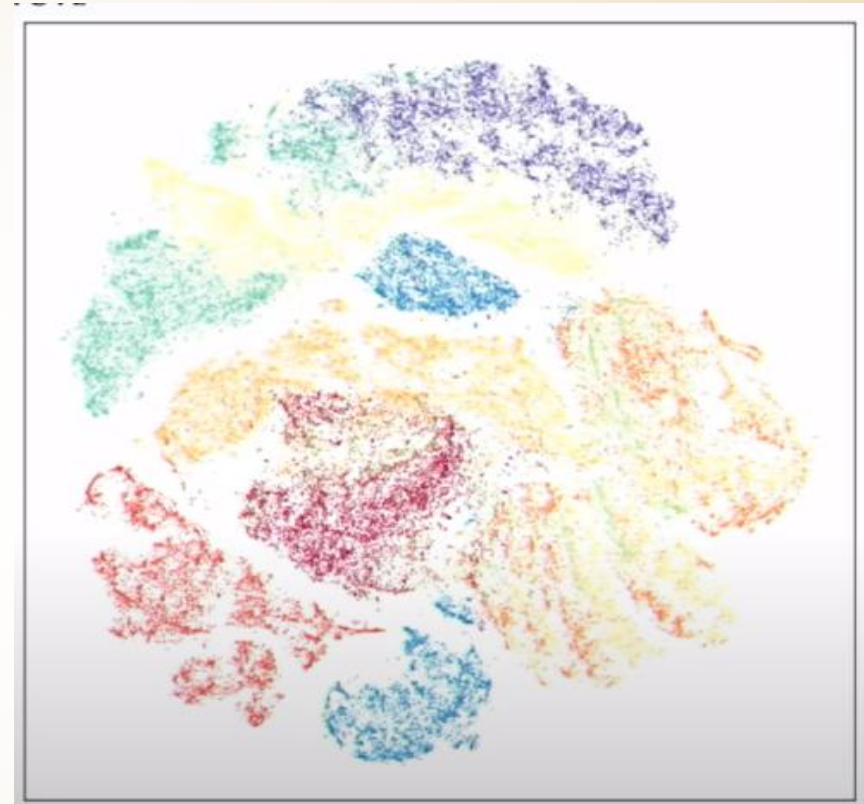


# Exemple MNIST



## UMAP

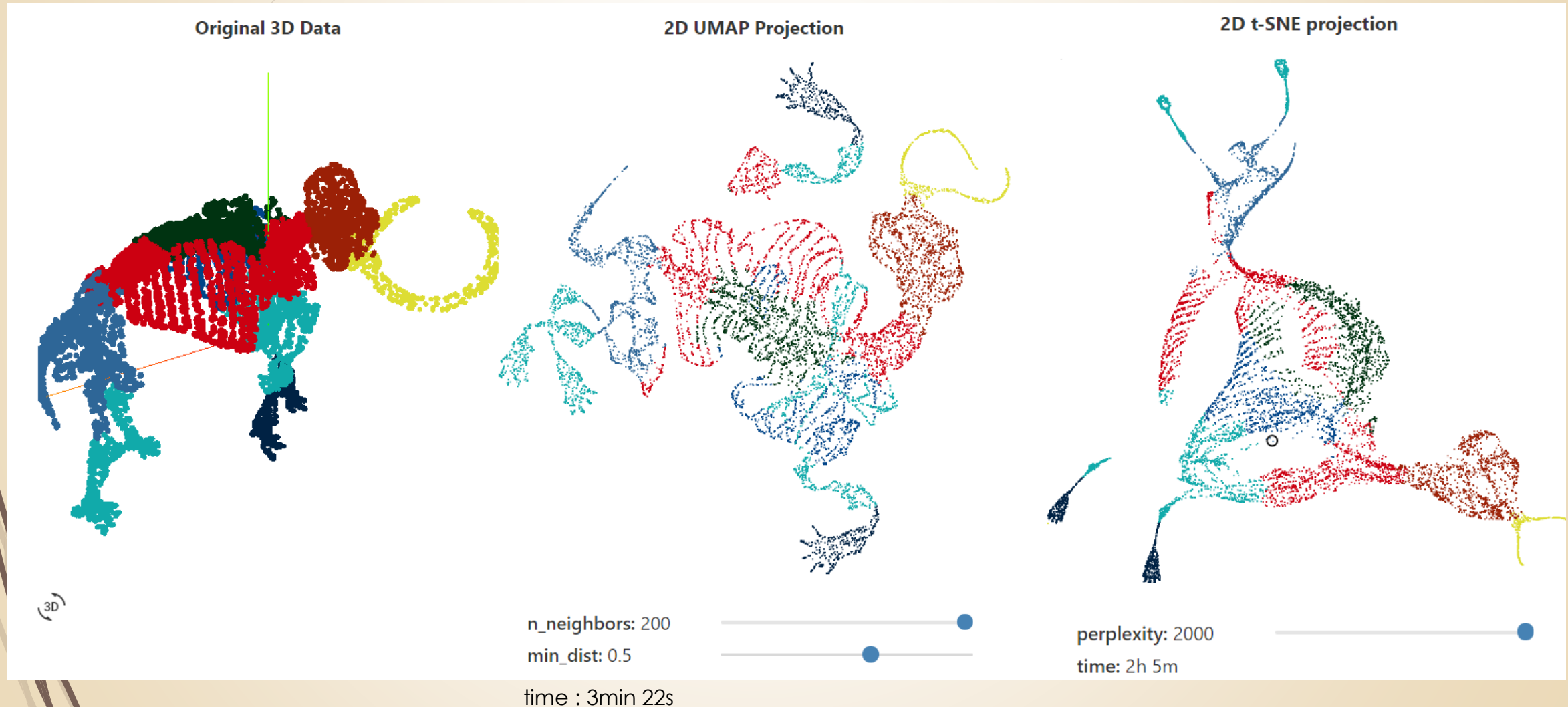
- Bonne séparation des groupes
- Conservation de la structure globale



## T-SNE

- Bonne séparation des groupes

# Comparaison des algorithmes en image





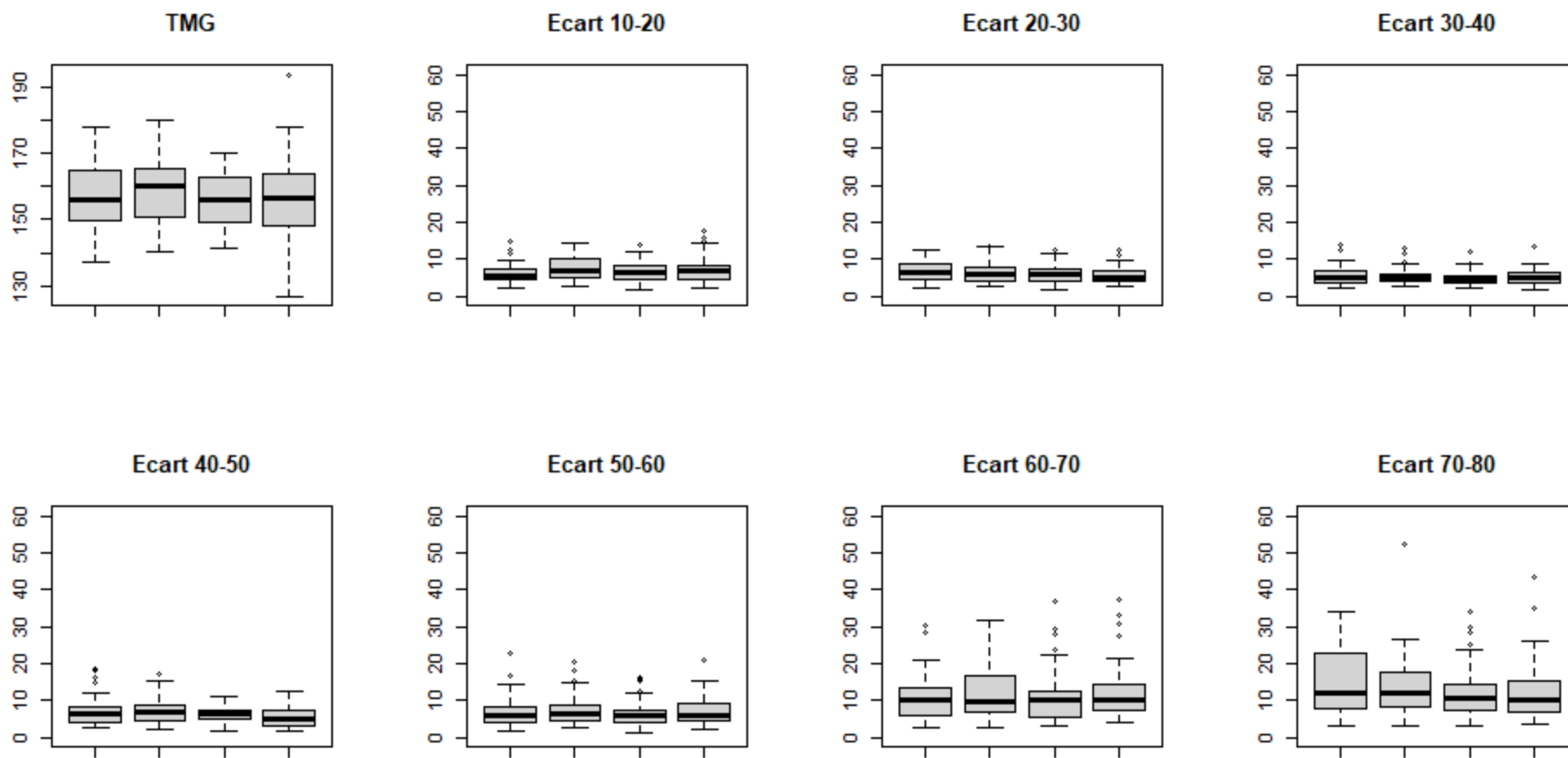


# Explication des données

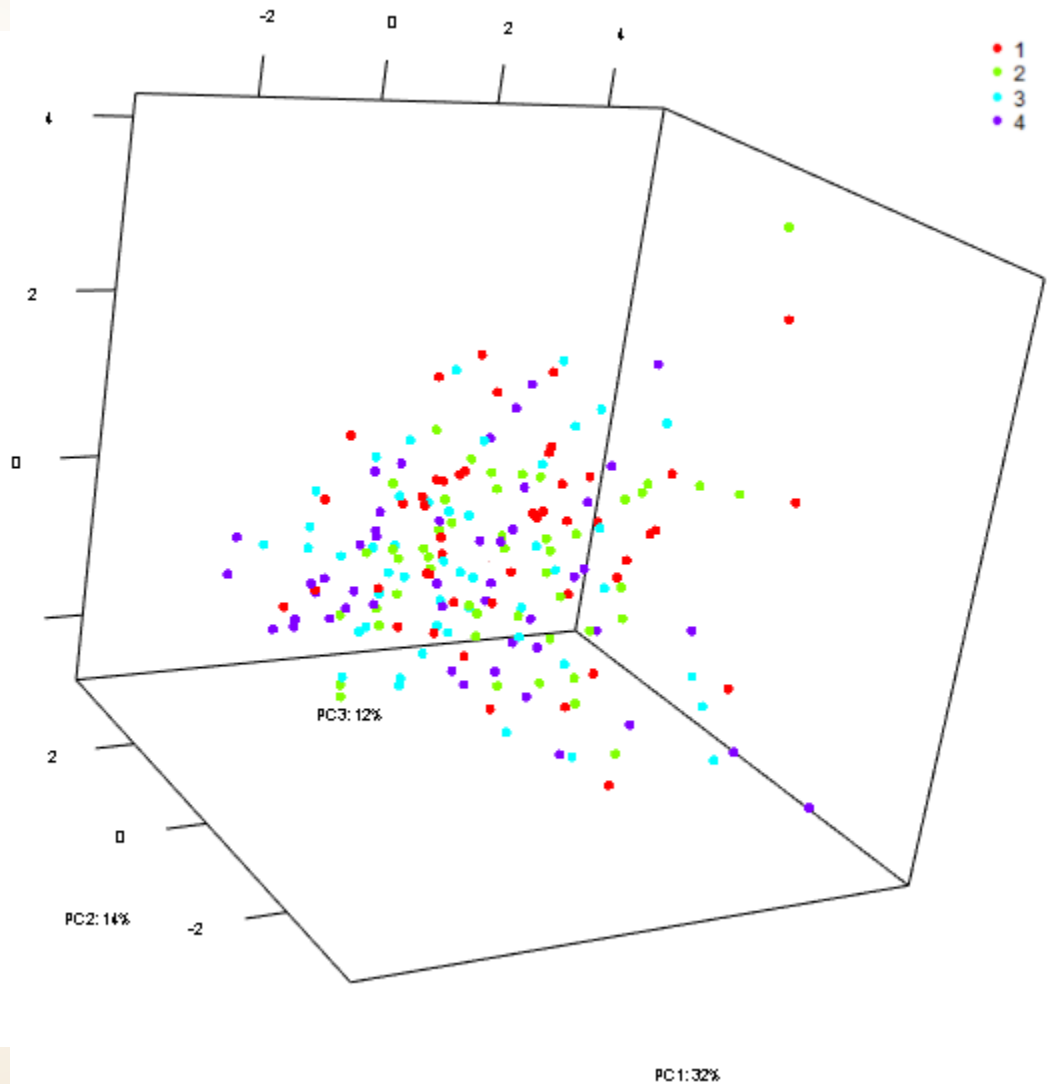
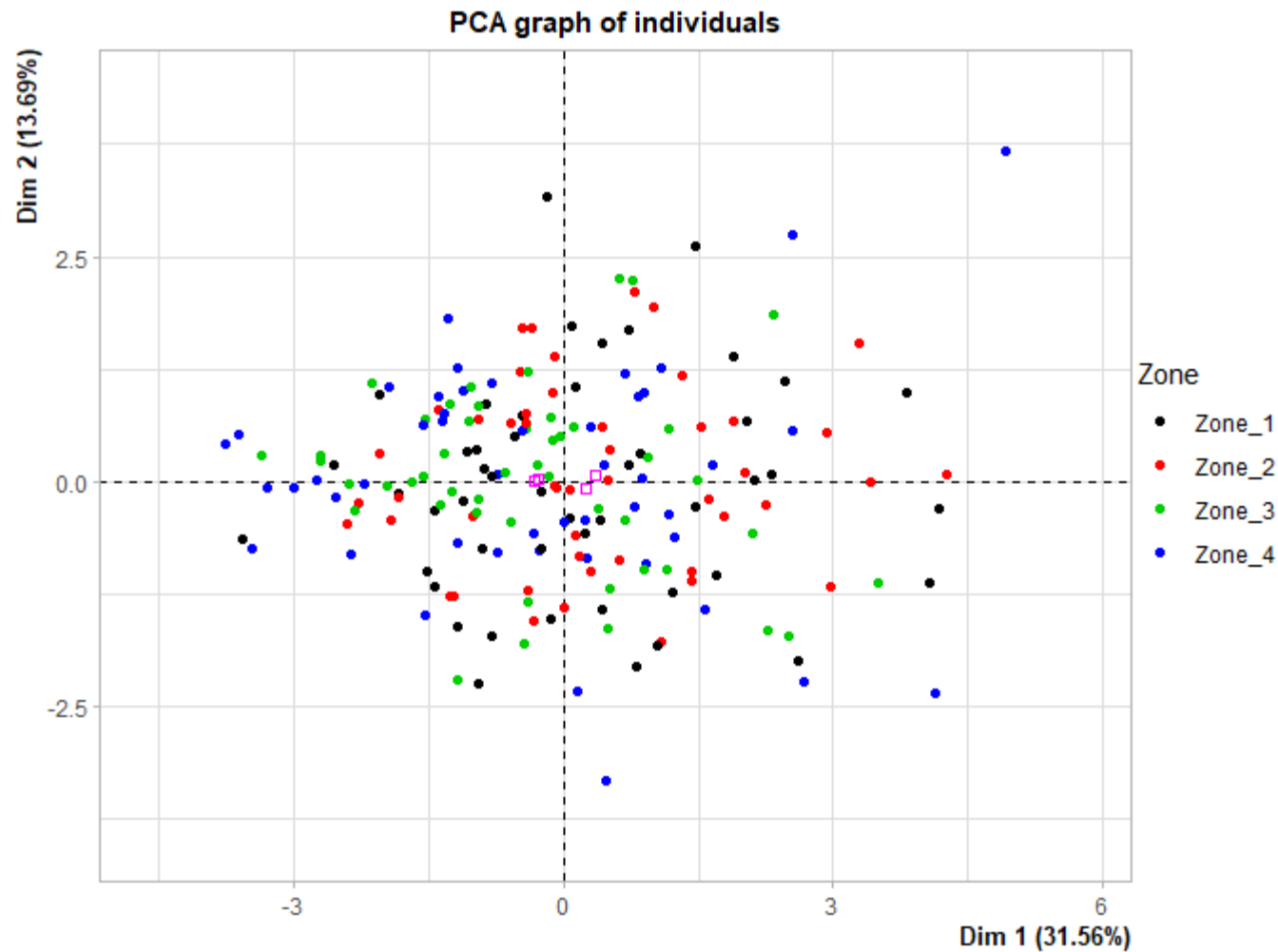
- Suppression de la dernière colonne (T90) car trop peu de données (~30% manquantes)
- Suppression des individus ne pouvant pas être modifiés (NaN) grâce à leur(s) répétition(s)
- Moyennisation des répétitions

# Représentation simple des données

Boxplot des vitesses de germination en fonction de la zone

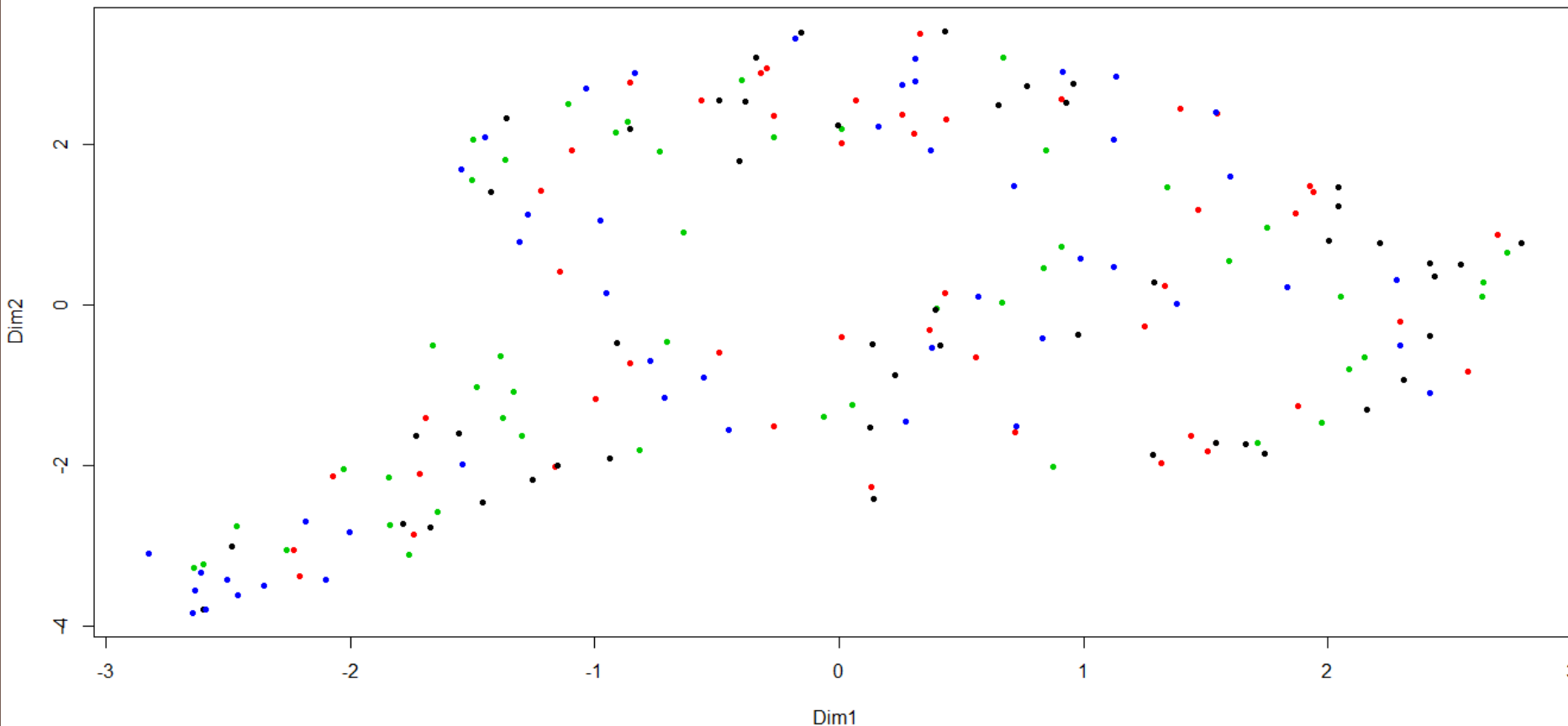


# ACP : recherche d'effet de la zone



# UMAP (Uniform Manifold Approximation and Projection)

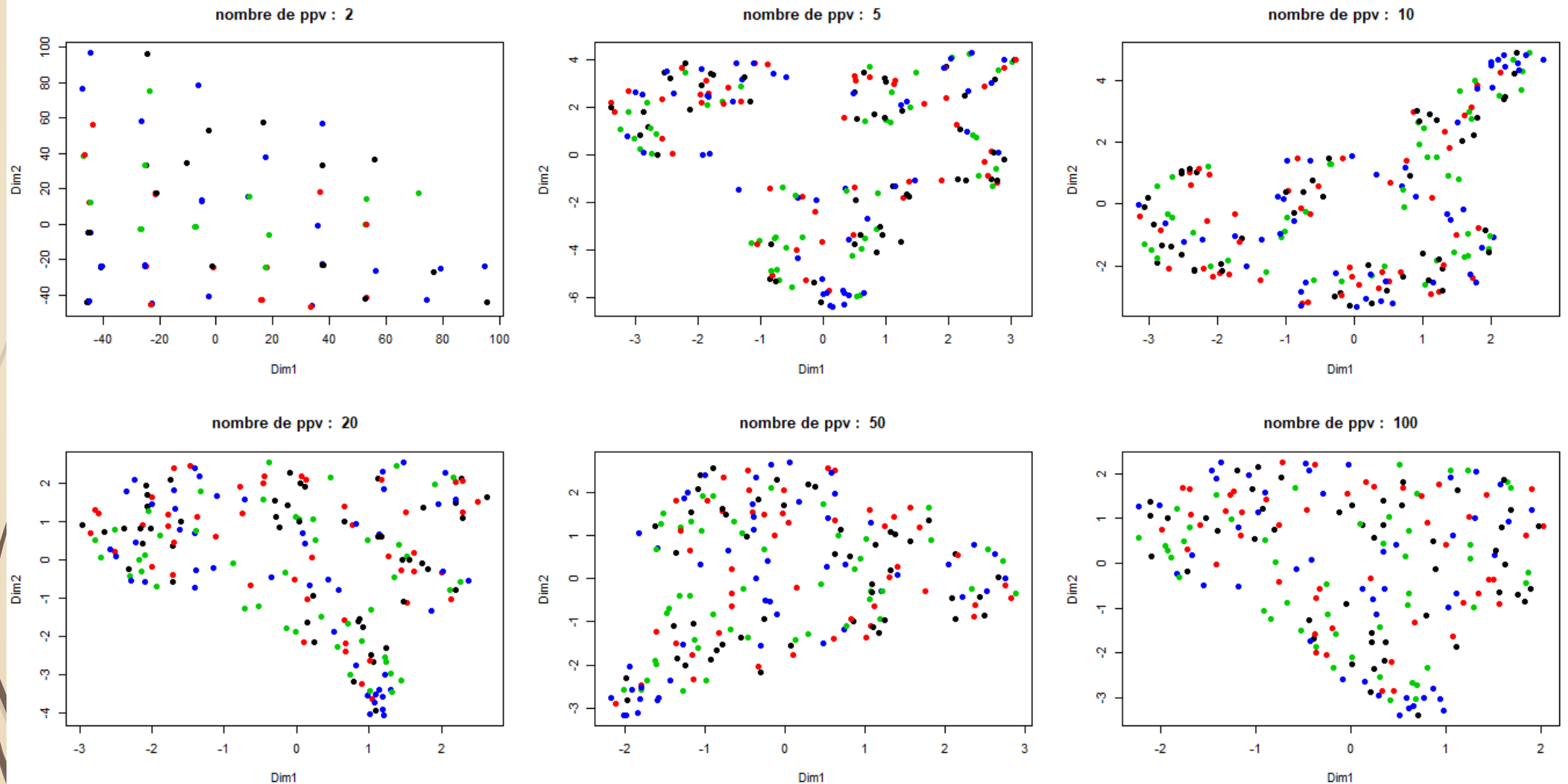
Paramètres par défaut



Paramètres par défaut :

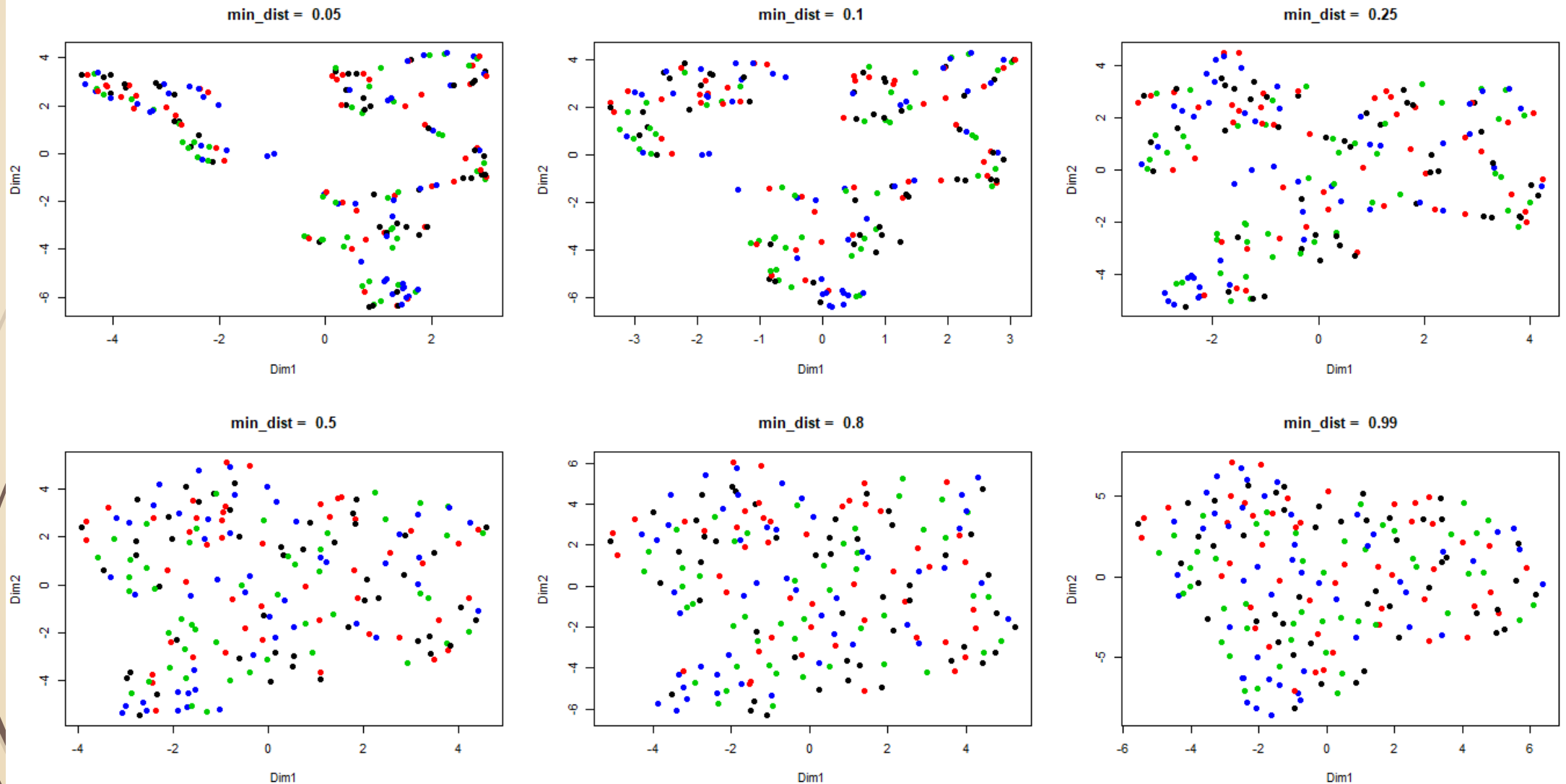
- 15 ppv
- Métrique euclidienne
- Distance minimum 0,1

# Variation du paramètre du nombre de plus proches voisins

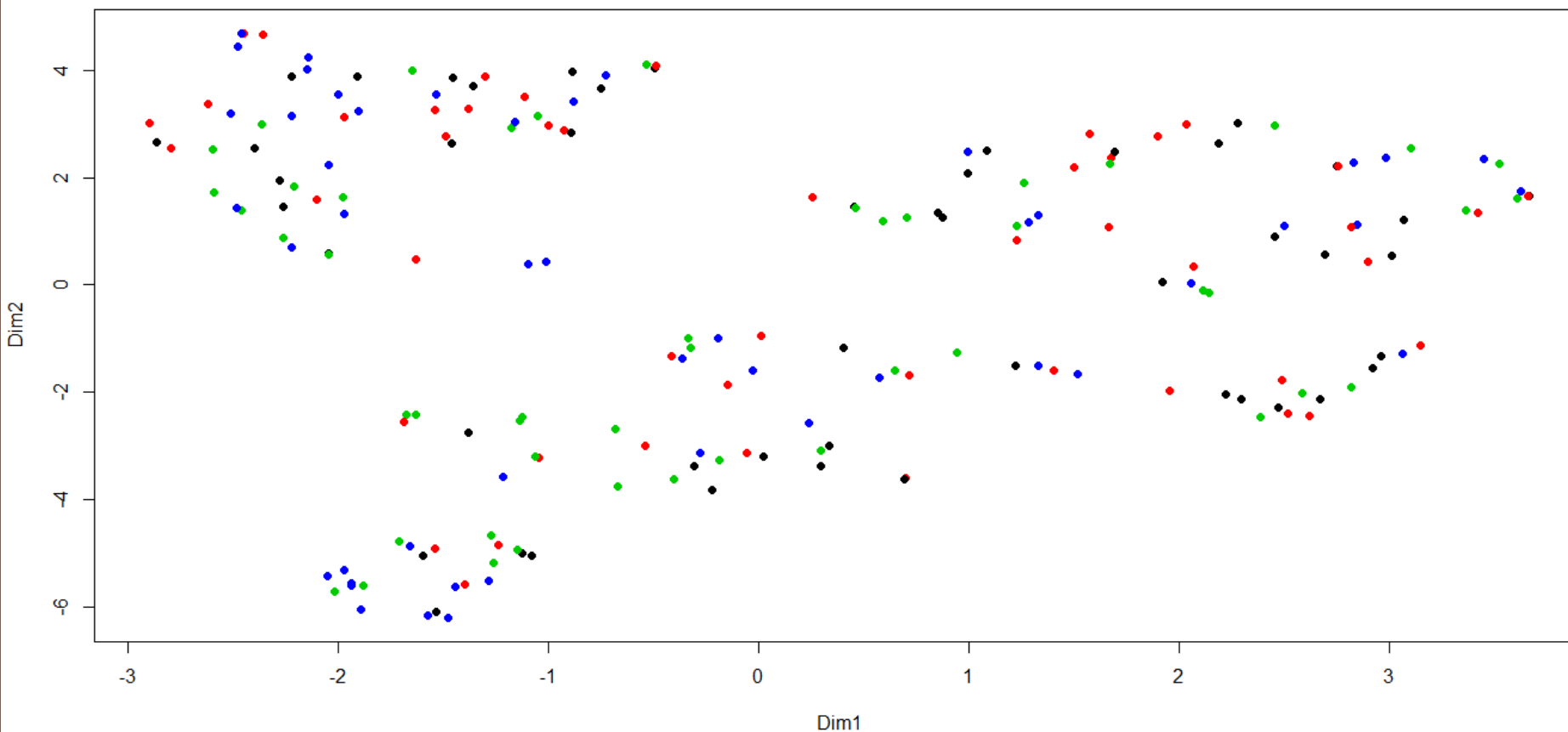




# Variation du paramètre de la distance minimum



# Choix final des paramètres



Paramètres choisis :

- 5 ppv
- Métrique euclidienne
- Distance minimum 0,1

# Conclusion

Aucun effet dû à la zone

Avantages :

- plus rapide que t-SNE
- respect de la structure (globale et locale)

Inconvénients :

- pas de signification des axes
- ne sépare pas 2 clusters imbriqués



# Sources

- <https://pair-code.github.io/understanding-umap/>
- Video explicative (anglais) :  
<https://www.youtube.com/watch?v=nq6iPZVUxZU>
- [https://umap-learn.readthedocs.io/en/latest/how\\_umap\\_works.html](https://umap-learn.readthedocs.io/en/latest/how_umap_works.html)