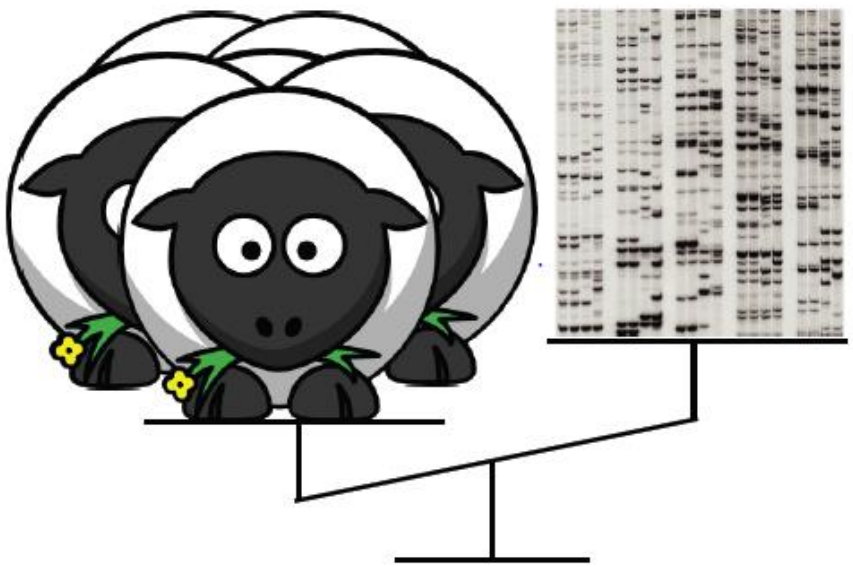
Multivariate environmental characterization of samples

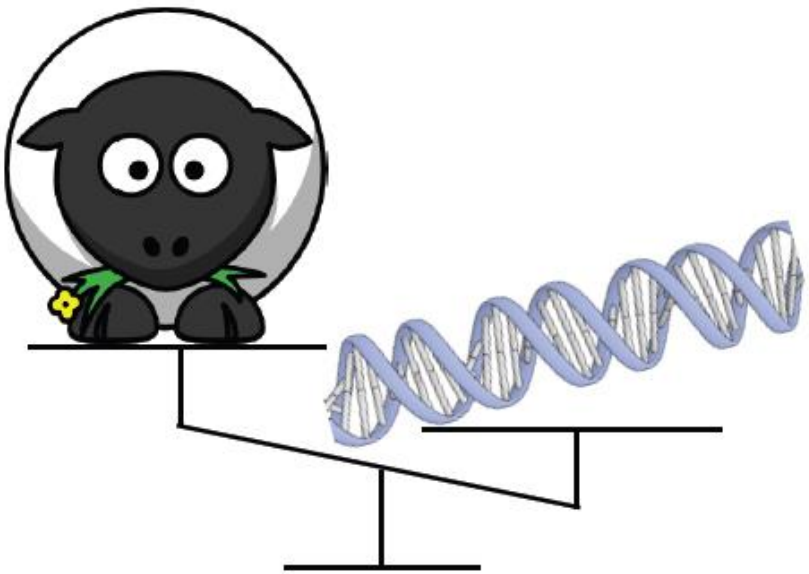
Dr Stéphane Joost – Oliver Selmoni (Msc)
Laboratory of Geographic Information Systems (LASIG)
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Maximising environmental contrast

- Landscape genomics best works when environmental contrast is maximized between sampling locations
- This is the main reason to identify optimal sampling locations
- There is another reason, another constraint
- The recent availability of whole genome sequence (WGS) data implies to reconsider sampling strategies in landscape genomics for economic reasons
- While we had many individuals and few genetic markers ten years ago, we now face the contrary with high costs of WGS limiting the number of samples to be sequenced
- Molecular resolution is excellent but it is achieved at the expense of spatial representativeness and statistic robustness (we must remove sampling points)
- Therefore, starting from a standard sampling, it becomes comnpulsory to apply sub-sampling strategies in order to keep most of the environmental variation



(a) Avant : beaucoup d'individus et peu de marqueurs génétiques.



(b) Après : beaucoup de marqueurs génétiques et peu d'individus.

Moroccan example (Nextgen)

- To study local adaptation of goats and sheep's breeds in Morocco, we used a sampling design based on a regular grid across the territory
- In each cell of this grid, 3 individuals were sampled in 3 different farms.
- Then, the final subset selected for sequencing had to meet two criteria in order to ensure a regular cover of both physical (x,y) and environmental (z) spaces
  1. using a stratified sampling technique over a range of climatic variables previously filtered by a PCA
  2. minimizing a clustering index in order to ensure spatial spread
- In Morocco, the sub-sampling procedure using a hierarchical clustering resulted in two datasets: 162 goats selected over 1283, and 162 sheep over 1412
- Based on variables such as temperature, pluviometry and solar radiation. By maximising the environmental information collected, we were able to select individuals that are the most relevant to study adaptation.

Approach

- Individuals must be continuously distributed along environmental gradients to regularly sample the space of climatic parameters
- Taking into account climatic extremes does not provide the adequate information to calibrate generalized linear models and is not useful for population genetics.
- We use PCA to transform K correlated variables (environmental variables) into K uncorrelated factors, the main components

The role of PCA

- The first axis represents the direction along which the points vary the most
- The following axes explain in turn the maximum variance of the remaining variance, while being orthogonal to each other
- The components are ordered according to the variance they explain

The role of PCA

- The PCA on environmental variables provides a synthetic picture of the climate
- It allows to choose samples "as different as possible": the first main axes explain most of the climatic variability while limiting random noise
- The Euclidean distance in this subspace makes it possible to define an ecological distance between the farms
- This limits the weight given to highly correlated factors when selecting indivuals to sample

Hierarchical Ascendant Classification (HAC)

- The choice of a limited number of points representing habitat diversity involves grouping farms according to their similarities (based on ecological distances) and selecting one for each group
- We used a HAC with the Ward criterion
- Iterative method where the two closest elements (points and/or groups) are gathered at each stage
- Ward's algorithm chooses which elements to merge by minimizing the increase in intra-class inertia
- This criterion tends to bring together close classes containing few elements
- HAC is deterministic and the partitions generated at each iteration are nested. This is why the process is often represented in the form of a tree.

Constrained draw (selection of points)

- The sampling strategy based on a grid provides a regular density of points on the territory
- The random choice of one farm per class is likely to introduce a spatial bias
- This is why the spatial representativeness of the selected samples is quantified with a distribution index (D)
- We define D as the sum of the distances between each point and its nearest neighbor in the selection
- The higher D, the more distant the points are from each other and the better their spatial distribution

Sheep in Morocco - PCA

- 432 farms with sheep
- 117 environmental variables
- We kept 7 first components (96% of variance)
- The space generated by these axes reliably represents the climatic conditions

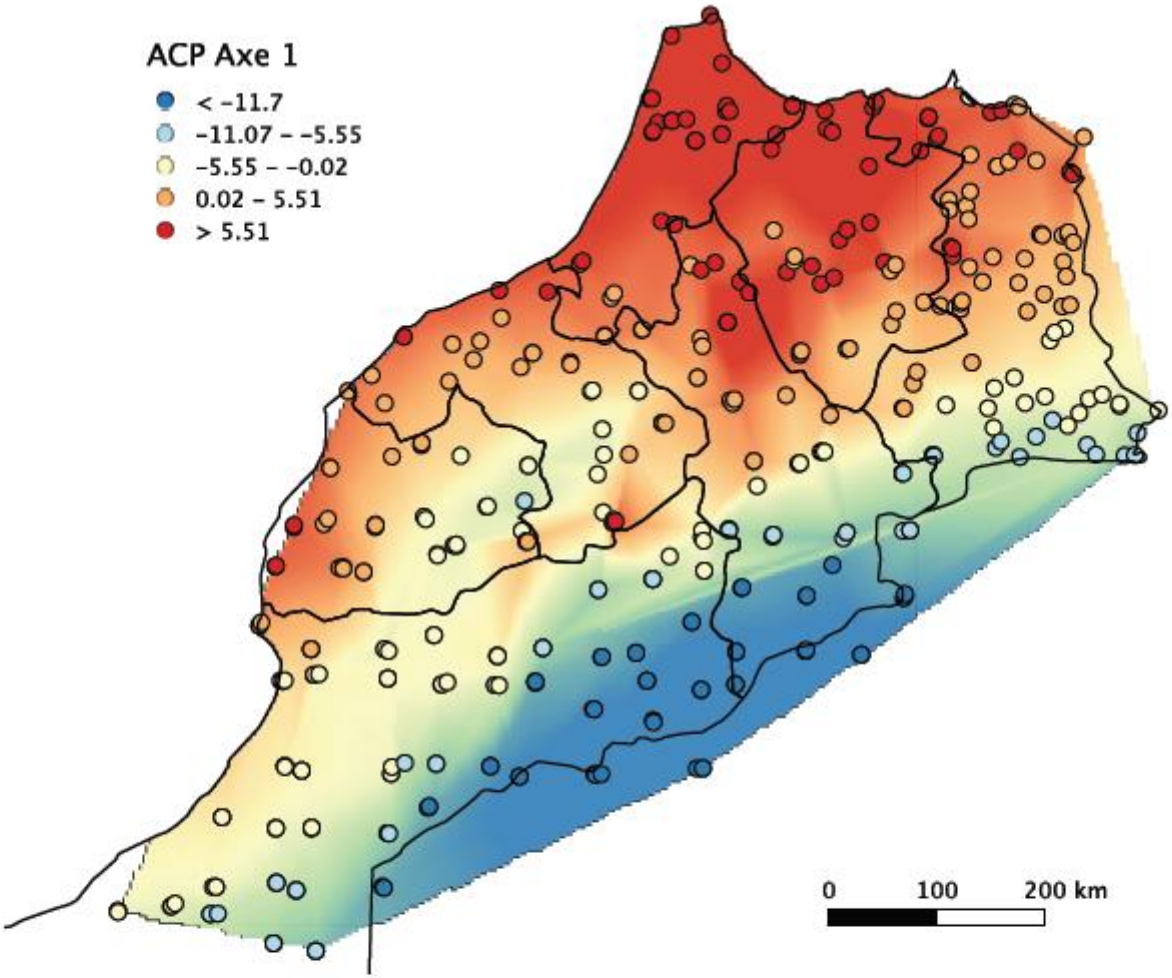


Figure 4.4 – La première composante principale distingue les régions côtières des régions désertiques du Maroc. Les valeurs élevées (en rouge) sont corrélées avec le nombre de jours de pluie tandis que les valeurs basses (en bleu) se situent près du désert où l'ensoleillement est important et où les précipitations varient peu. L'interpolation des valeurs est présentée en arrière-plan pour faciliter la visualisation de leur distribution spatiale. Modèle numérique d'altitude : SRTM3.

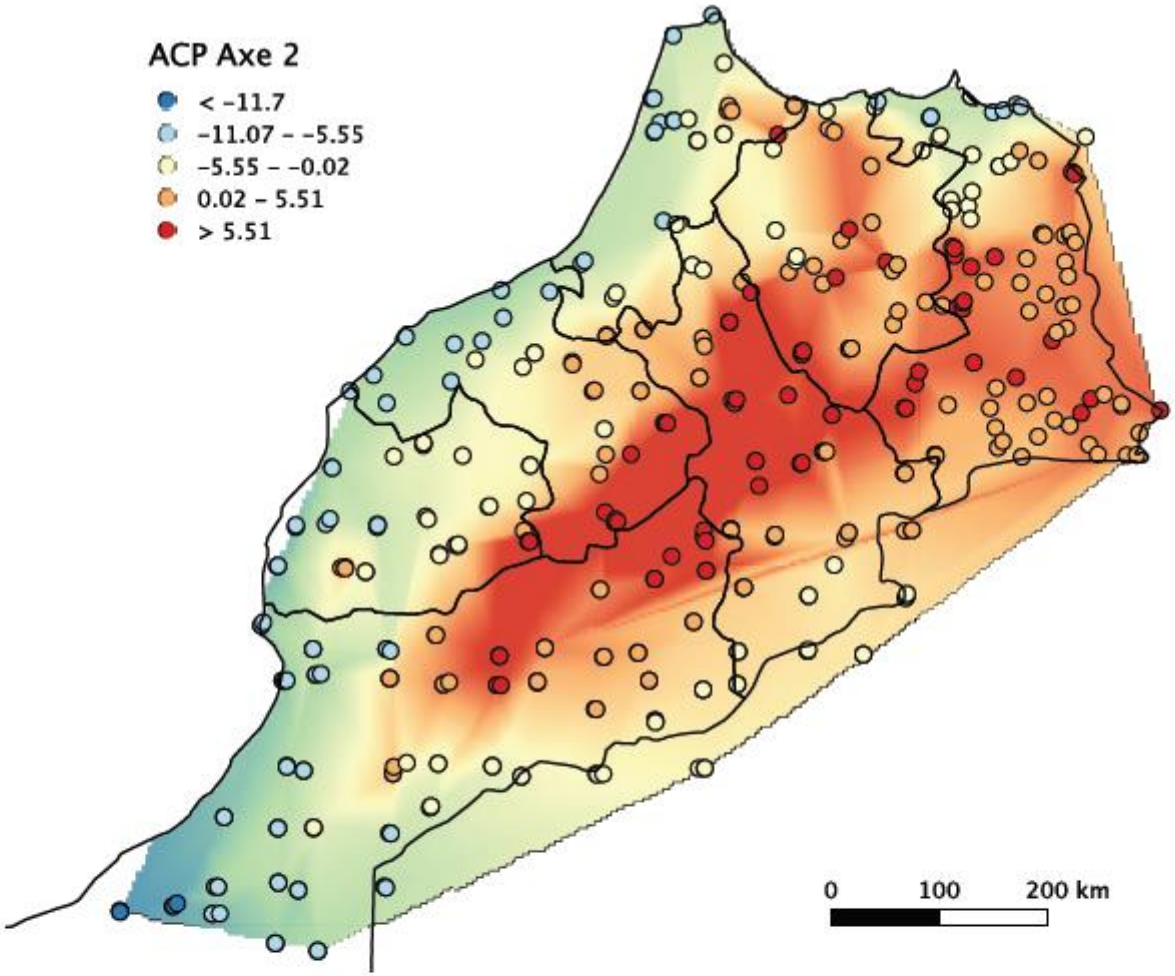
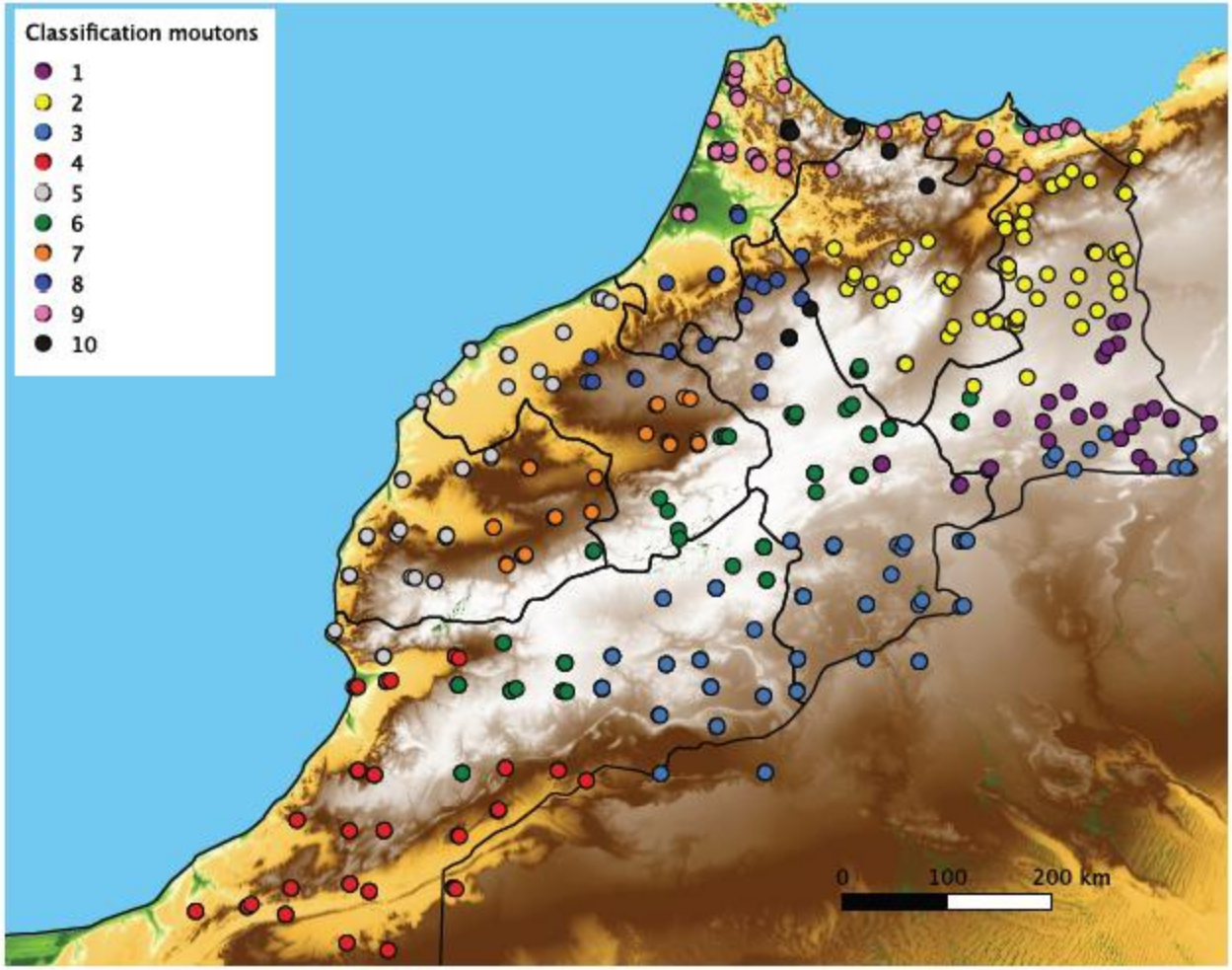


Figure 4.5 – La deuxième composante principale est corrélée avec l'altitude. Les valeurs élevées (en rouge) se situent dans la chaîne de l'Atlas où la température hivernale est basse, alors que l'amplitude de température diurne, le nombre de jours de gel et les précipitations y sont élevés. L'interpolation des valeurs est présentée en arrière-plan pour faciliter la visualisation de leur distribution spatiale. Modèle numérique d'altitude : SRTM3.

Sheep in Morocco - HAC

- The Euclidean distance in this space makes it possible to define an environmental distance between the farms with the perspective of their aggregation
- Some farms are in the same climate cell and have been grouped for analysis, which leads to 229 different climatic conditions for sheep
- These groups of farms represent distinct points in the climatic conditions and served as basic units for HAC
- The following map shows the classification for 10 classes



(b) Fermes élevant des moutons

Figure 4.6 – Positions des fermes élevant des petits ruminants. Les couleurs représentent la classification en fonction des conditions climatiques. Comme escompté, les fermes sont globalement regroupées en fonction de leur proximité spatiale et de l'altitude. La classification distingue également les régions côtières de l'intérieur des terres. Modèle numérique d'altitude : SRTM3.

HAC in 164 classes

- As we have money for WGS for 164 individuals, we must apply a HAC to get 164 classes
- Each class contains 1-3 farms = 1-18 individuals
- Finally 50 random draws of one individual per class and retained the one with the highest spatial distribution index (D)

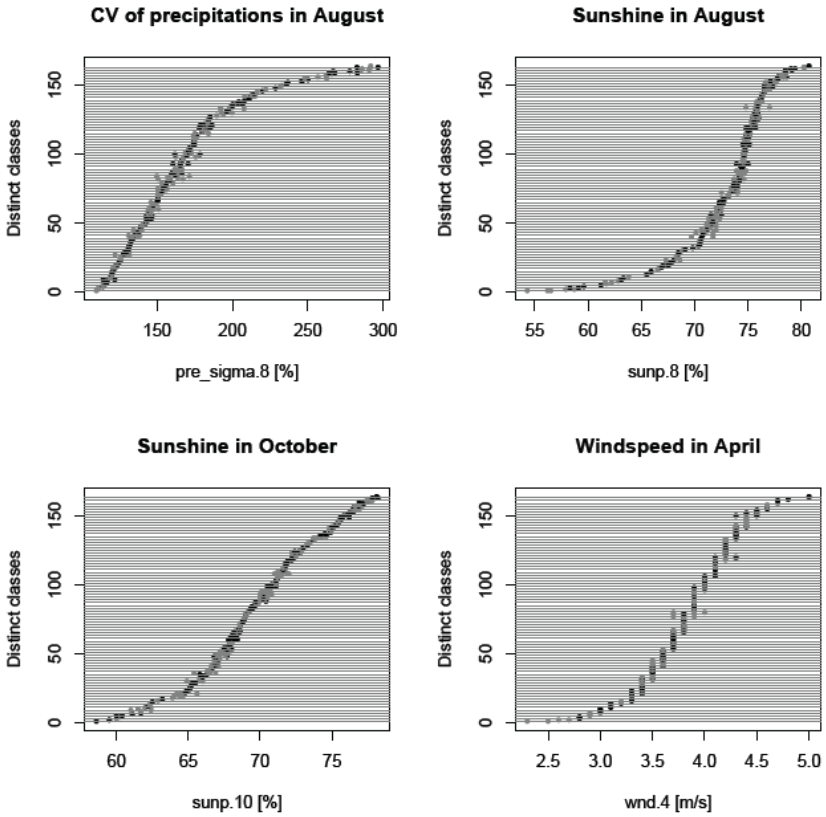


Figure 4.7 – Les 164 classes sont représentées en fonction de quatre variables environnementales d'origine. Chaque ligne montre les valeurs prises par les fermes appartenant à une classe. Par souci de lisibilité, les classes ont été triées en fonction de la valeur moyenne des fermes.

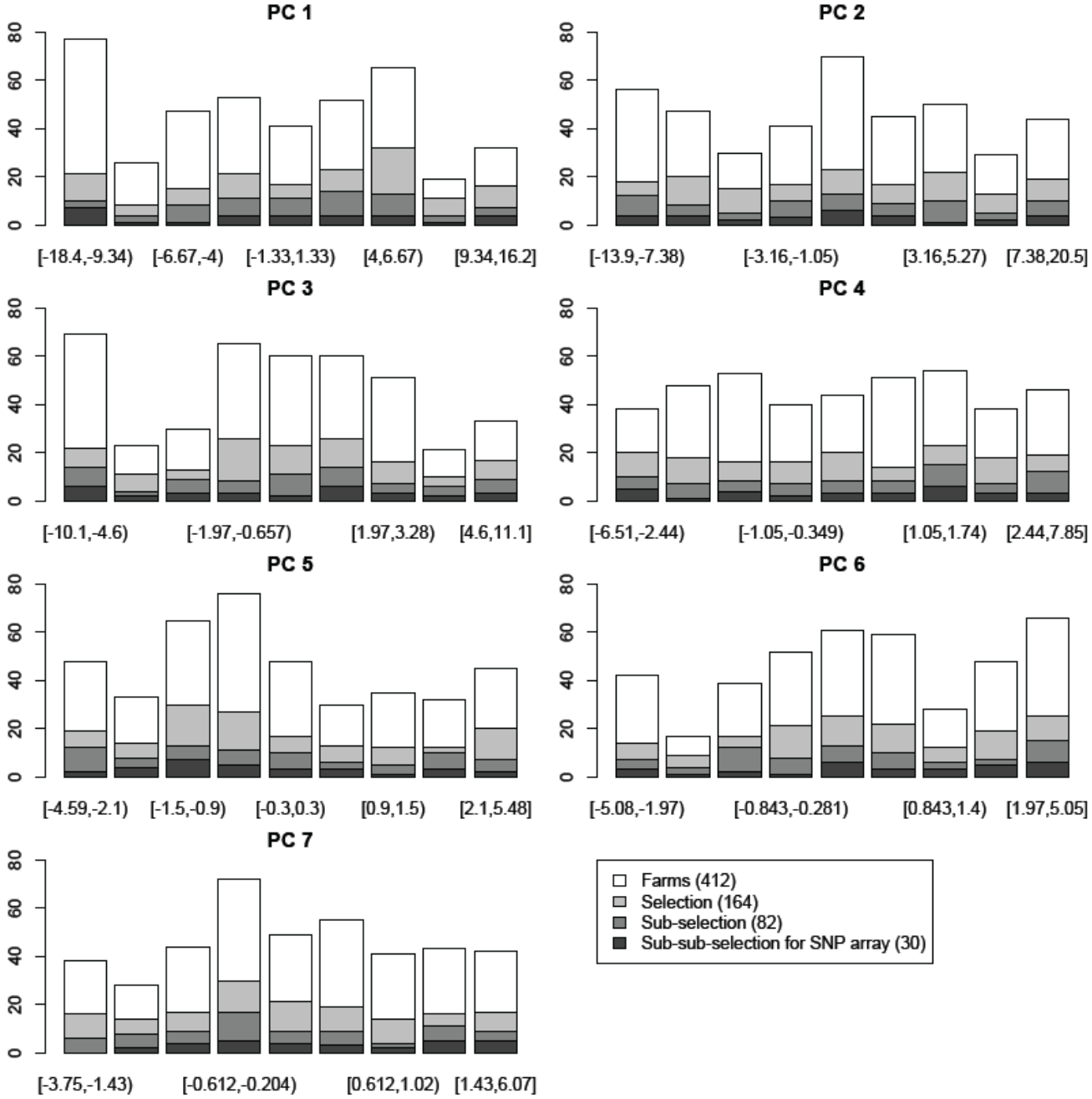


Figure 4.10 – Distribution des fermes de chèvres sur les axes principaux. Les trois ensembles d'individus sélectionnées sont imbriqués.

Exercise

- Characterizing samping points with multivariate environmental data