# Atelier de formation sur les réseaux bayésiens dans le cadre de l'ANR Floodscale

Sandra Perez

# Atelier de formation sur les réseaux bayésiens dans le cadre de l'ANR Floodscale

Sandra PEREZ

Il existe sur le marché plusieurs logiciels de réseaux bayesiens. Le choix s'est porté sur BayesiaLab en raison à la fois de sa puissance et de sa simplicité d'utilisation.

## 1. Importer les données

Bayasialab peut lire des données issues de fichiers texte ou csv ou bien des données issues d'une base.


Les données que nous allons utiliser sont des données horaires issues de SAFRAN pour l'année 1999-2000. Elles correspondent pour les 23 BV d'étude de Floodscale aux :

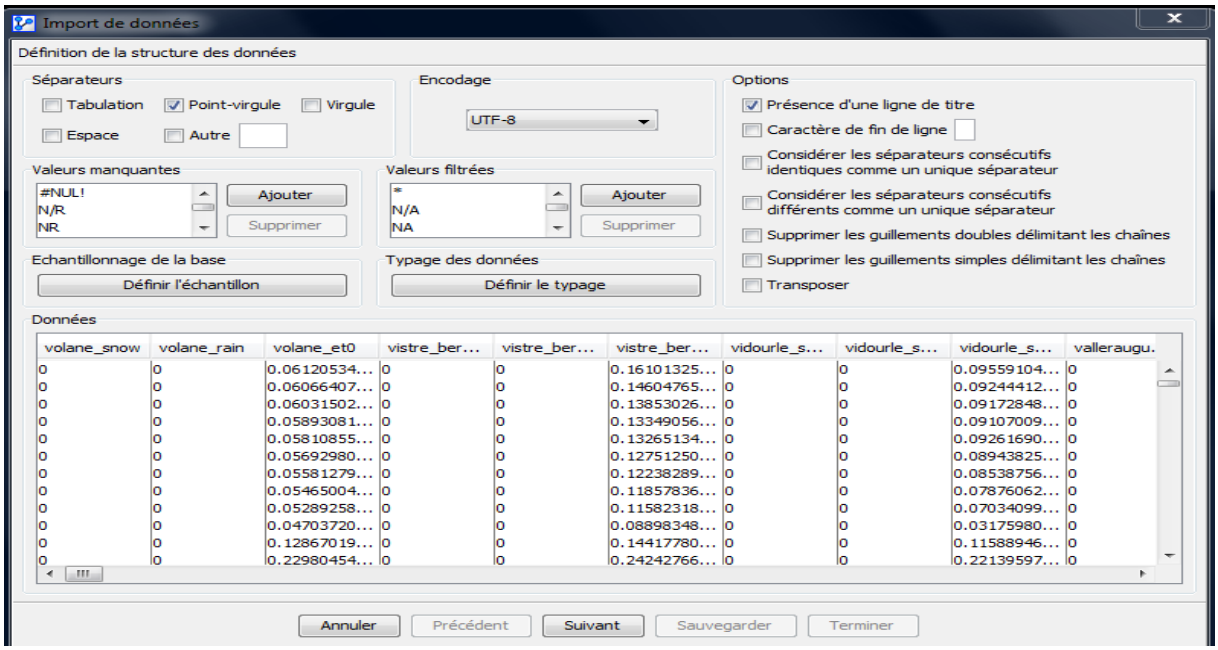
-**données de météo** : pluies liquides (rain mm/h), aux pluies solides (snow en mm/h), et à l'évapotranspiration de référence (ETo en mm/h)

-**données de débits** : mm/h

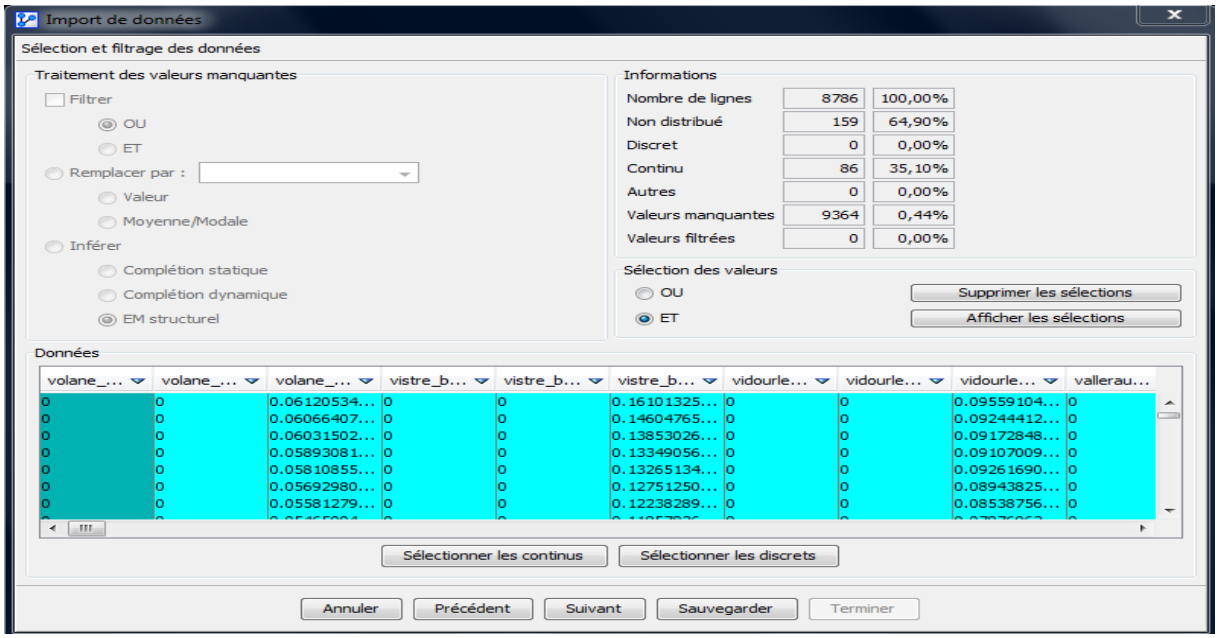
-**caractéristiques des bassins versants** :

| | | Area $A$ (km²) | Average slope $\beta$ (°) | Total length $L$ of stream | Length $L_{perm}$ of permanent streams (km) | Dominant geology | Dominant soil texture | Mean elevation (m) | % of forest cover | % of agricultural areas cover |
|---|---|---|---|---|---|---|---|---|---|---|
| #1 | Alzon (Uzès) | 71,0 | 3,9 | 66,9 | 29,2 | 1 | 1 | 157 | 37,0 | 48,9 |
| #2 | Ardèche (Meyras) | 98,7 | 22,3 | 167,4 | 130,1 | 2 | 2 | 899 | 58,7 | 3,8 |
| #3 | Arre (Le Vigan) | 155,0 | 18,4 | 213,0 | 95,7 | 3 | 1 | 660 | 67,1 | 7,9 |
| #4 | Auzonnet (Les Mages) | 49,0 | 14,9 | 66,6 | 17,6 | 4 | 3 | 360 | 60,5 | 9,2 |
| #5 | Beaume (Rosières) | 200,0 | 21,0 | 340,0 | 259,2 | 3 | 2 | 653 | 69,0 | 6,9 |
| #6 | Cloutasses | 0,4 | 8,5 | 0,7 | 0,2 | 5 | 2 | 1410 | 4,6 | 0,0 |
| #7 | Dardaillon (St-Just) | 36,4 | 2,1 | 48,5 | 20,3 | 6 | 3 | 38 | 2,4 | 89,5 |
| #8 | Dourbie (Dourbies) | 42,9 | 12,0 | 58,1 | 30,6 | 5 | 2 | 1227 | 68,4 | 1,7 |
| #9 | Gagnière (Gagnières) | 55,3 | 18,4 | 93,9 | 59,3 | 3 | 2 | 545 | 77,1 | 1,2 |
| #10 | Gardon (Saumane) | 104,0 | 22,9 | 145,7 | 66,2 | 3 | 1 | 682 | 59,9 | 1,1 |
| #11 | Gardon de Sainte-Croix | 47,0 | 18,4 | 50,8 | 20,6 | 3 | 1 | 773 | 60,4 | 1,7 |
| #12 | Gardon de Saint-Germain | 30,5 | 20,8 | 46,0 | 15,1 | 3 | 1 | 667 | 66,8 | 3,0 |
| #13 | Gardon de Saint-Martin | 30,5 | 21,2 | 46,2 | 17,5 | 3 | 1 | 622 | 67,6 | 0,7 |
| #14 | Herault (Valleraugue) | 46,2 | 25,7 | 78,1 | 28,5 | 3 | 1 | 856 | 79,0 | 1,3 |
| #15 | Latte | 0,2 | 10,5 | 0,4 | 0,1 | 5 | 2 | 1393 | 44,8 | 0,0 |
| #16 | Lez (Montferrier-sur-Lez) | 115,0 | 7,0 | 66,9 | 21,4 | 1 | 4 | 144 | 17,6 | 32,0 |
| #17 | Rieumalet (Pont-de-Montvert) | 20,0 | 13,4 | 25,0 | 24,4 | 5 | 2 | 1341 | 41,5 | 2,0 |
| #18 | Salaison (Mauguio) | 50,8 | 3,5 | 60,0 | 16,9 | 4 | 4 | 78 | 11,9 | 51,8 |
| #19 | Tarn (Pont-de-Monvert) | 67,0 | 11,1 | 92,4 | 82,1 | 5 | 2 | 1296 | 46,8 | 0,5 |
| #20 | Valescure | 3,9 | 23,8 | 7,2 | 3,2 | 5 | 2 | 513 | 100,0 | 0,0 |
| #21 | Vidourle (Sauve) | 190,0 | 10,1 | 281,7 | 55,3 | 4 | 4 | 271 | 41,1 | 19,2 |
| #22 | Vistre (Bernis) | 291,0 | 2,4 | 219,6 | 64,1 | 6 | 3 | 79 | 3,0 | 56,4 |
| #23 | Volane (Vals-les-Bains) | 109,0 | 20,8 | 148,6 | 77,6 | 7 | 2 | 818 | 64,5 | 1,8 |

elles sont contenues dans le fichier RB9900




Une première fenêtre indique au logiciel la structure des données : séparateur, type de valeurs manquantes (possibilité d'en ajouter), possibilité de transposer les données, présence d'une ligne de titre ou non[1].




Un clic droit sur la flèche du bas indique les valeurs de certains paramètres classiques : min, max, moyenne, ecart type etc.

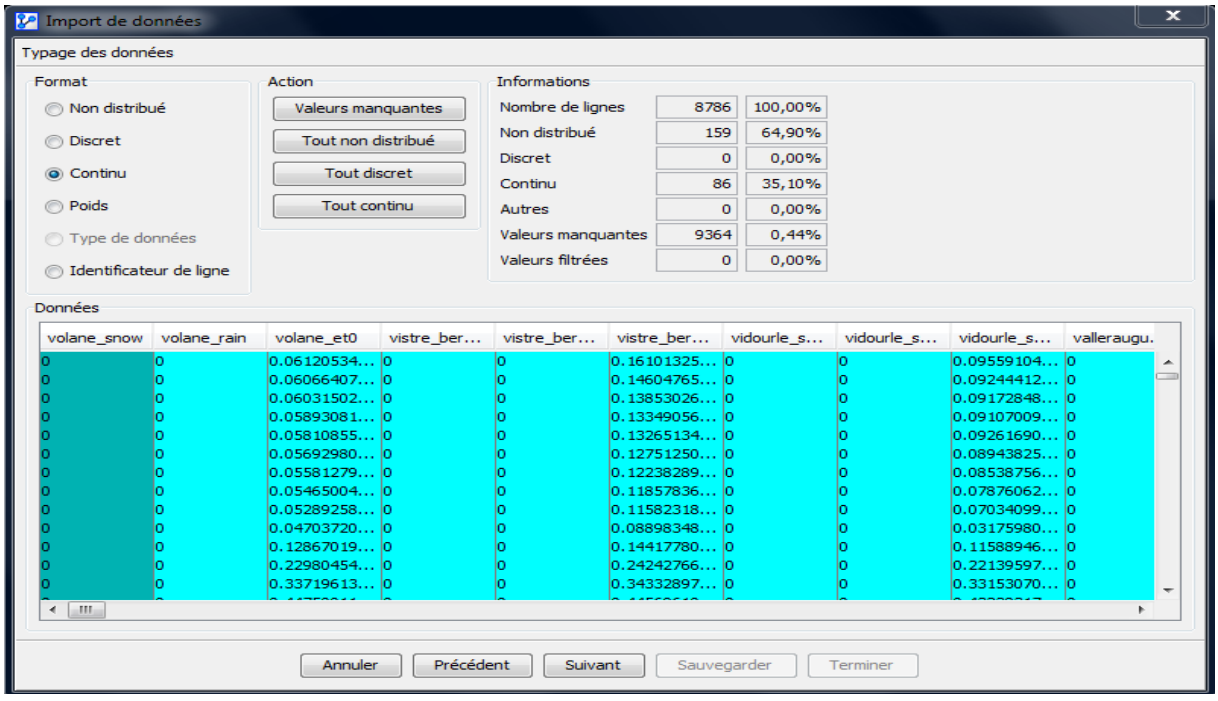
[1] Il est également possible de filtrer les données selon un critère qui parait pertinent (par exemple : valeur de débit horaire élevé) : cliquer sur l'icône de débits de volane par exemple, et on cliquerait une fois sur la modalité qui correspond aux crues pour quelle soit surlignée en bleu et définir un filtre Et (un second click passant en rouge indiquerait un filtre Ou)

## 2. Traitement des valeurs manquantes

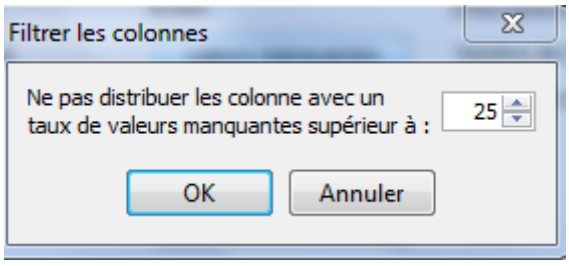
Bayesialab offre la possibilité de traiter les valeurs manquantes en leur attribuant par exemple une valeur spécifique choisie par l'utilisateur, ou bien la moyenne de la série si la variable est continue, ou modale si elle est discrète. On peut également utiliser la puissance de l'inférence qui peut être de 3 types : complétion statique (), complétion dynamique (à chaque modification structurelle du réseau tel que l'ajout, l'inversion, ou la suppression d'arc, les valeurs manquantes sont dynamiquement estimées par le réseau et les données disponibles et remplacées par les valeurs modales) ou bien par un algorithme de type EM structurel (*la probabilité de chaque modalité des variables à valeurs manquantes sont dynamiquement estimées avec la structure courante et des données disponibles. Ces probabilités sont directement utilisées pour l'apprentissage de la structure et de ses paramètres, c'est à dire qu'il n'y a pas de complétion par une modalité spécifique. Cette méthode est la plus précise mais aussi la plus gourmande en temps de calcul.* long)

nota : Le type de traitement choisi est propre à chaque variable possédant des valeurs manquantes. Si on souhaite appliquer le même traitement pour toutes ces variables, il faut sélectionner l'ensemble des colonnes (double click sur une colonne ou encore « CTRL + A »).

On a également un certain nombre d'informations sur les données (nombre de lignes, nombre de variables non distribuées, nombre de variables discretes ou continues etc...)




Action : Valeurs manquantes



Bayesialab fonctionne avec des variables de types discretes. Comme la plupart de nos variables sont continues nous devons les discrétiser (c'est la raison pour laquelle elles apparaissent comme non distribuées). Il est également possible de ne pas distribuer un champ si celui ci n'apporte pas d'information pertinente (exemple : un numéro d'identification des lignes).

Il existe plusieurs type de discrétisation possibles :

-Approximation de densité :

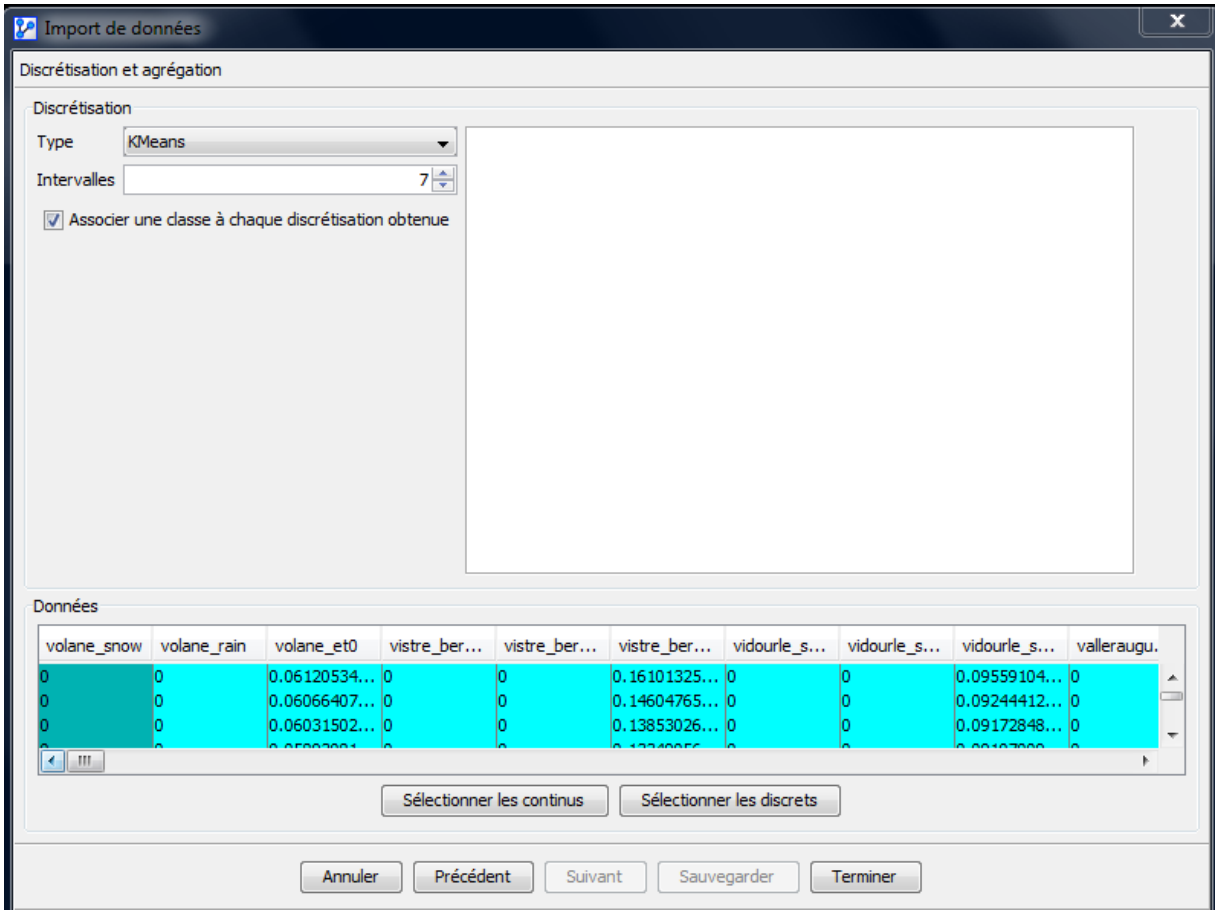
-Kmeans : chaque observation appartient à la partition avec la moyenne la plus proche

-Egale distance : le domaine de variation est découpé en intervalles de même longueur

-Egale distance normalisée : le domaine de variation est découpé en intervalles de même longueur puis transformation normale

-Egale fréquence : intervalles ayant le même nombre de cas associés

-Manuelle




En raison de la nature des données (données horaires) il n'y a pas de très grandes variations dans les séries, mais on a bcp de lignes (8786), on décide donc d'appliquer a priori une discrétisation Kmeans en 7 classes. Le logiciel nous indique que pour la première variable : volane snow une telle discrétisation n'est pas possible (bcp de 0 et peu de variations), il nous propose alors de la remplacer

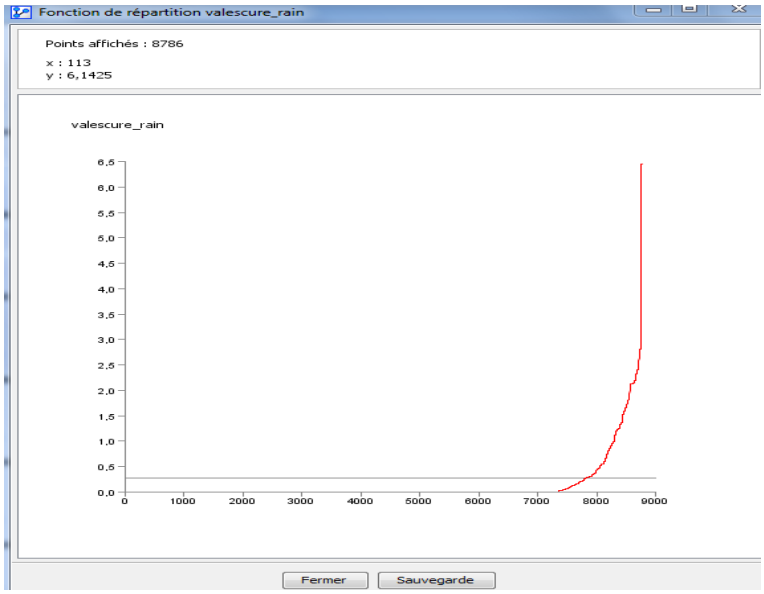
par une discrétisation de type "égale fréquence 2 classes", et de nous souvenir de ce choix pour des données qui seraient similaires.

Un rapport d'importation synthétise pour les différentes variables utilisées les modalités obtenues, leurs intervalles, et le type de discrétisation que l'utilisateur a demandé et celle retenue par le logiciel.

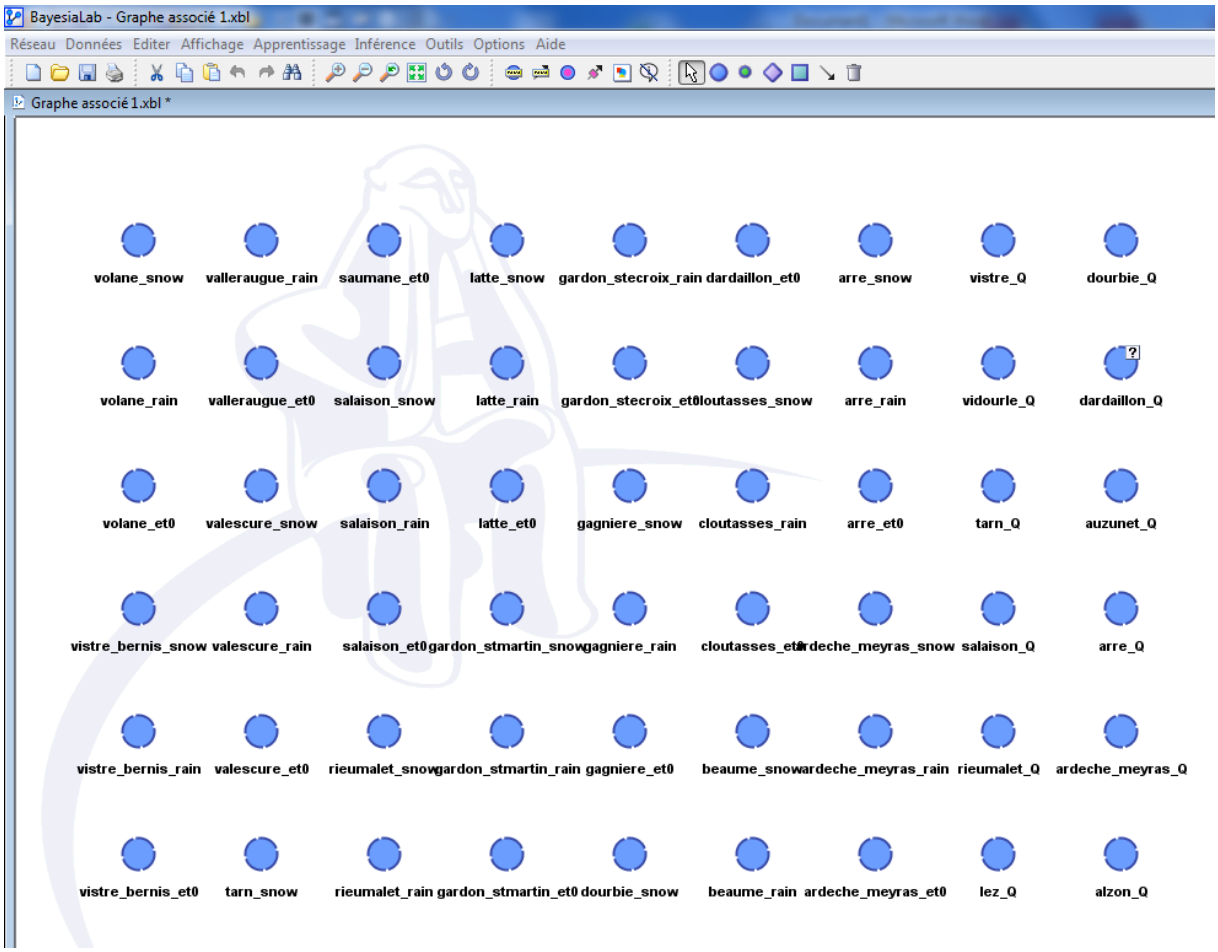
Rapport d'importation

| Noeuds 86 | | | | | |
|---|---|---|---|---|---|
| volane_snow | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0 | 0.0 | 0.0 | Demandée : KMeans - 7<br>Obtenue : Egales fréquences - 2 |
| | | >0 | 0.0 | 1.8721976 | |
| volane_rain | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0,353 | 0.0 | 0.35258937 | Demandée : KMeans - 4<br>Obtenue : KMeans - 2 |
| | | >0,353 | 0.35258937 | 5.6880364 | |
| volane_et0 | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0,075 | 1.45719E-4 | 0.07525587 | Demandée : KMeans - 4<br>Obtenue : KMeans - 4 |
| | | <=0,207 | 0.07525587 | 0.20741297 | |
| | | <=0,396 | 0.20741297 | 0.39598456 | |
| | | >0,396 | 0.39598456 | 0.70275533 | |
| vistre_bernis_snow | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0 | 0.0 | 0.0 | Demandée : KMeans - 4<br>Obtenue : Egales fréquences - 2 |
| | | >0 | 0.0 | 2.8787866 | |
| vistre_bernis_rain | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0 | 0.0 | 0.0 | Demandée : KMeans - 4<br>Obtenue : Egales fréquences - 2 |
| | | >0 | 0.0 | 3.6879296 | |
| vistre_bernis_et0 | Continu | Modalités | Intervalles | | Discrétisation |
| | | <=0,081 | 2.9829782E-4 | 0.08120633 | Demandée : KMeans - 4<br>Obtenue : KMeans - 4 |
| | | <=0,225 | 0.08120633 | 0.22452721 | |
| | | <=0,421 | 0.22452721 | 0.42102945 | |
| | | >0,421 | 0.42102945 | 0.7979687 | |

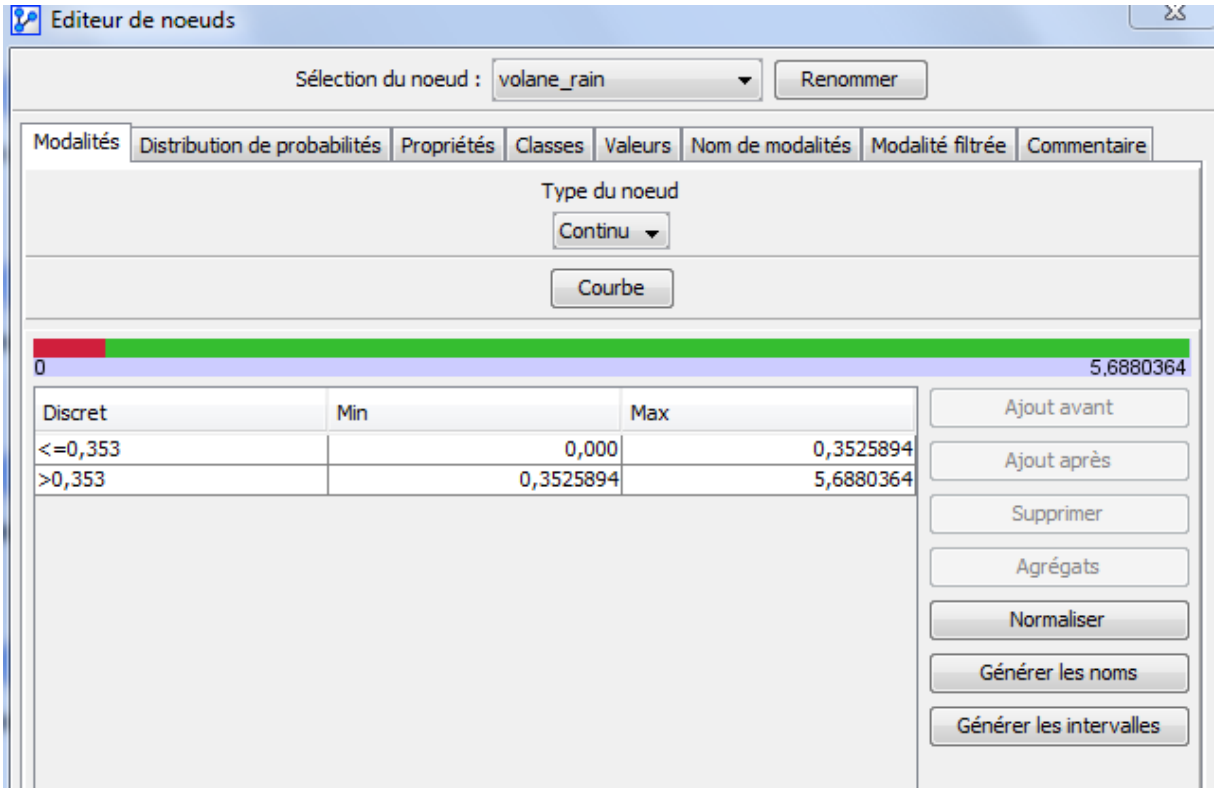
Dans le menu DONNEES l'utilisateur a la possibilité de réaliser des graphiques sur les variables telles que des fonctions de répartition :

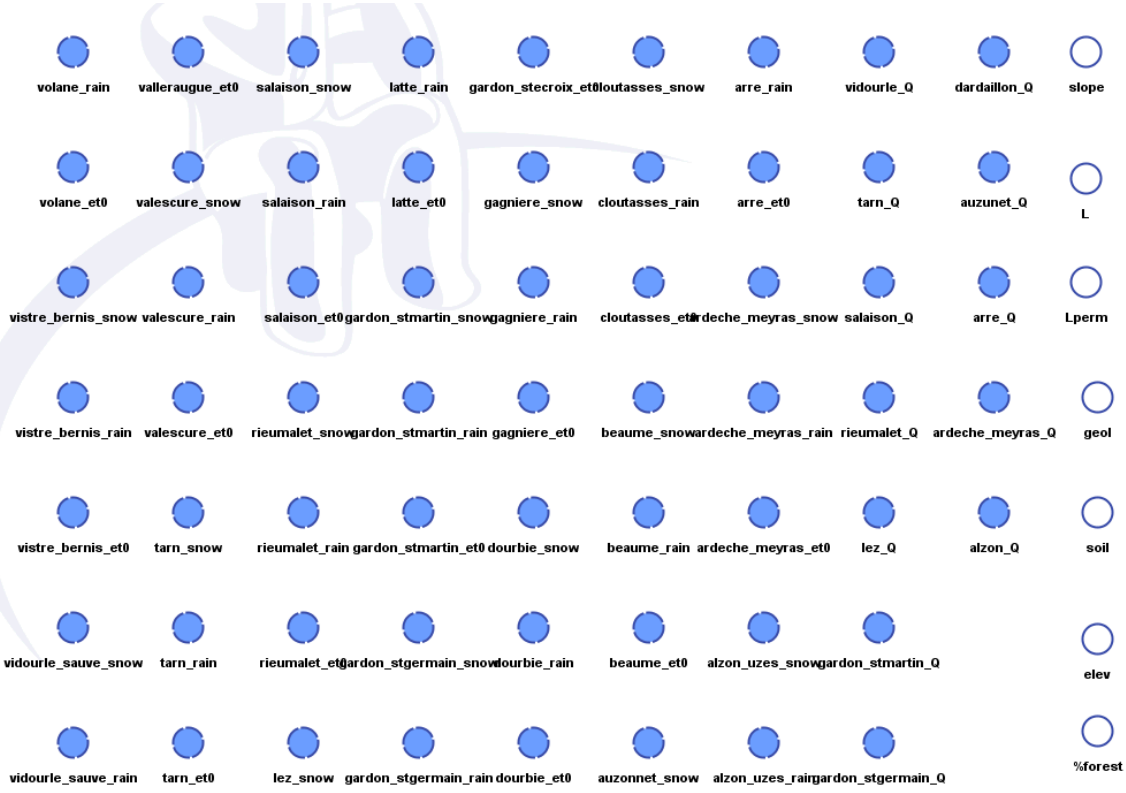


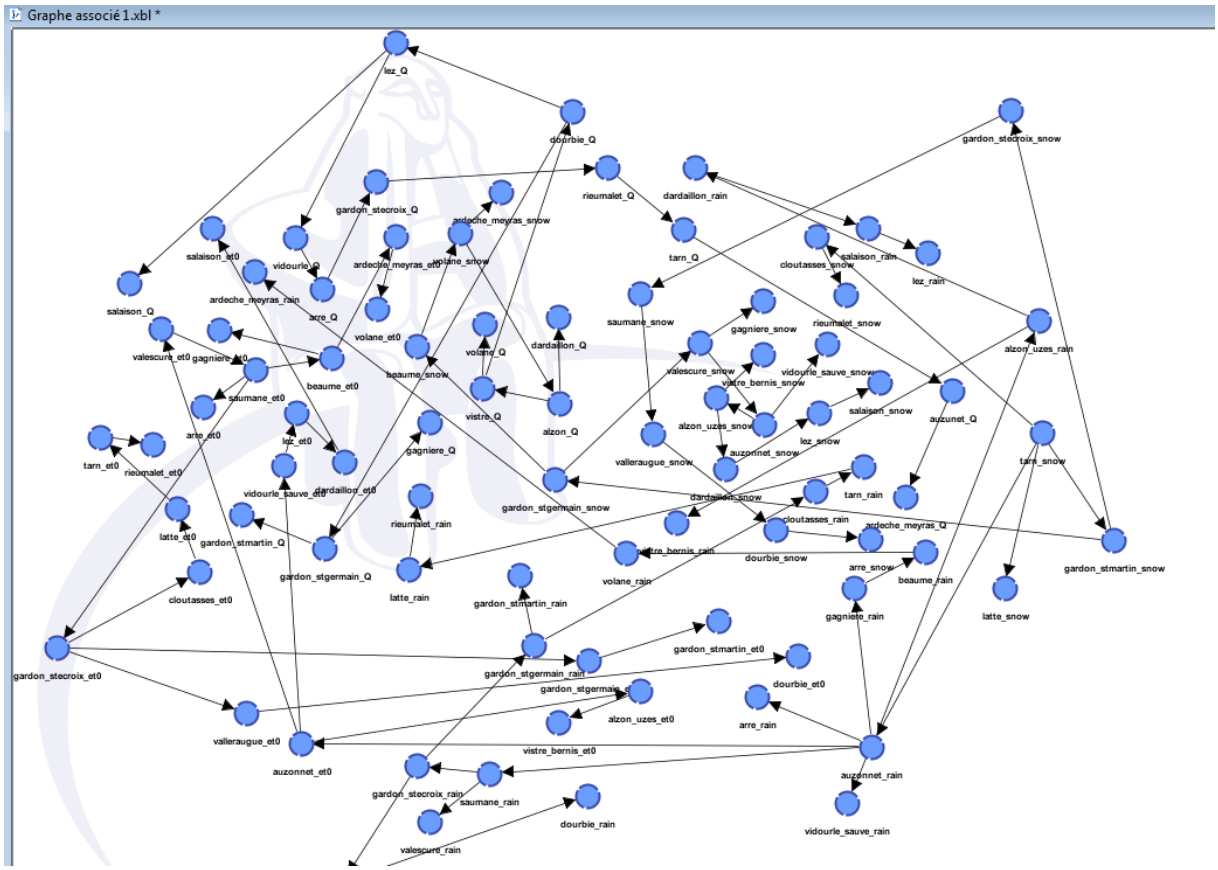

Après fermeture de l'assistant, une feuille de travail contenant les 86 nœuds correspondant aux différentes variables contenues dans la base de données apparaît.



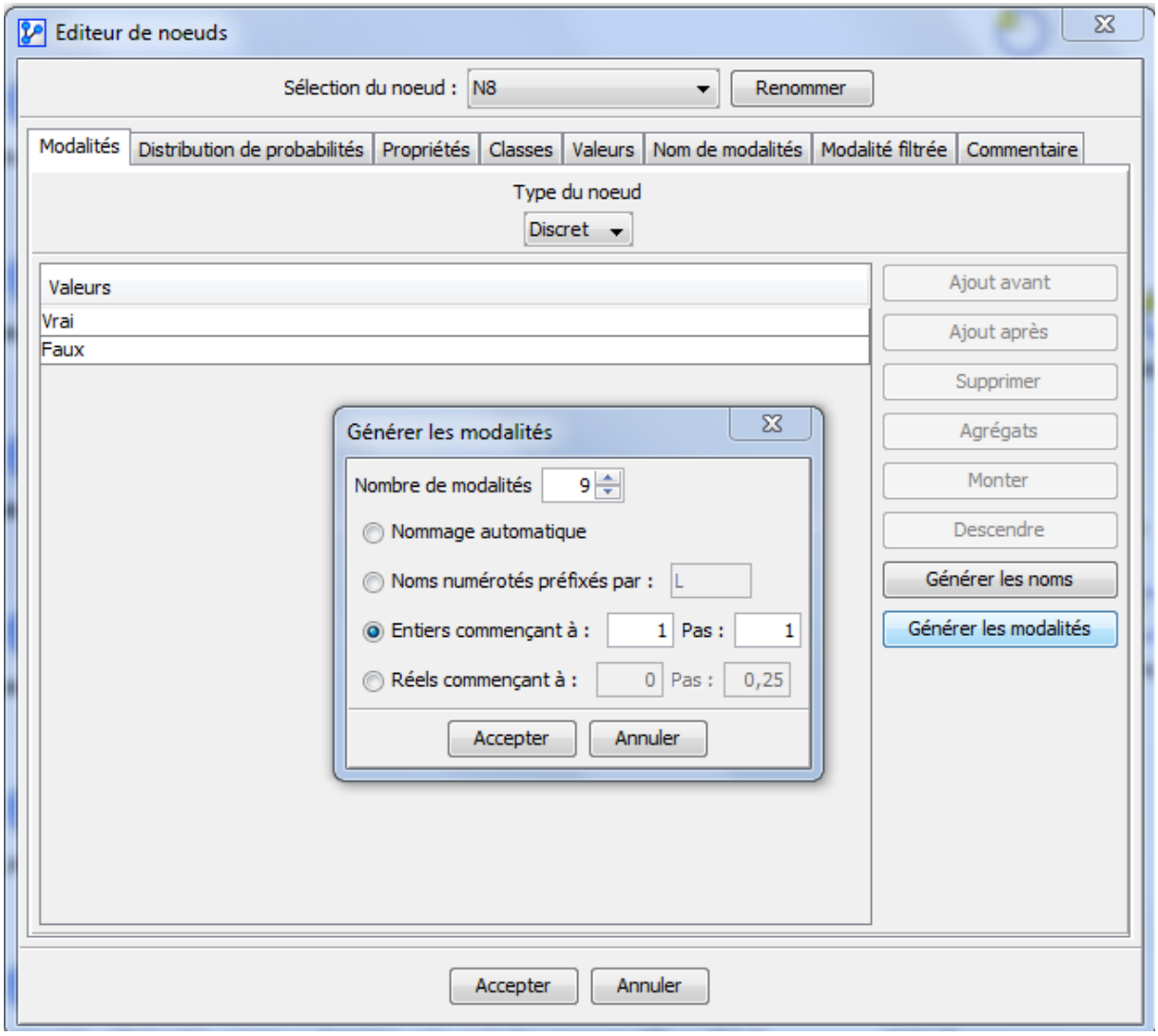
1 click droit sur un noeud, permet de l'éditer et d'obtenir un certain nombre d'informations :

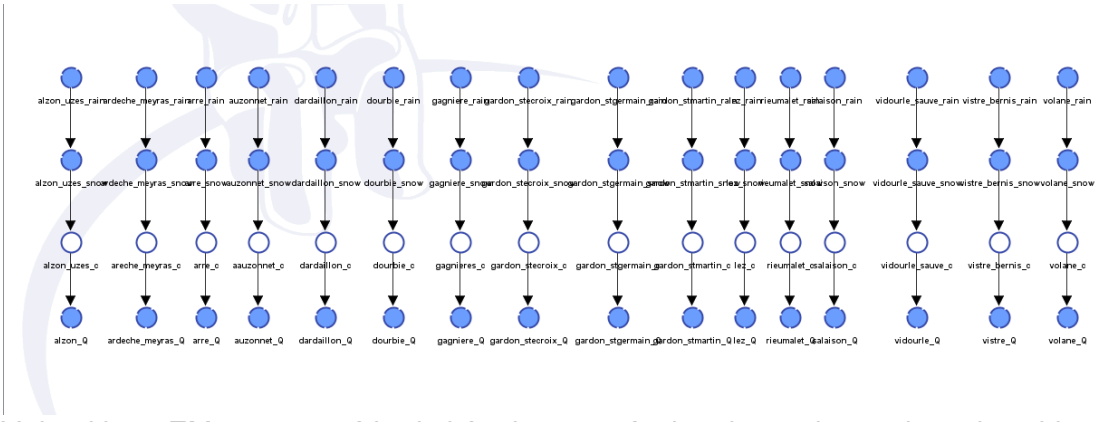





creer noeuds caracteristique apprentissage supervisé



L'algorithme EM est une méthode itérative pour évaluer le maximum de vraisemblance en présence de données incomplètes

### 2.2.4.1 Les chaînes de Markov

Dans le cadre d'un processus de Markov discret, le modèle le plus simple est la chaîne de Markov, encore appelée chaîne de Monte-Carlo. Soit $X_t$, la variable d'état au temps $t$, qui sont également les variables d'observations. $X_t$ peut prendre ses valeurs parmi un ensemble d'états discrets, $\{1, \cdots, N_e\}$. Pour une description probabiliste complète, il

faudrait prendre en considération tous les états précédant le temps $t$ pour évaluer la probabilité d'un état au temps $t$. Lorsque nous modélisons un processus par une chaîne de Markov, nous supposons que celle-ci ne dépend que de l'état précédent. Nous avons donc la simplification de la loi jointe suivante :

$$P(X_t = S_i | X_{t-1} = S_j, X_{t-2} = S_k, \cdots) = P(X_t = S_i | X_{t-1} = S_j) \quad (2.3)$$

Un tel processus peut être représenté par un schéma comme celui de la figure 2.7, dans le cas particulier ou $N_e = 3$.

1 2 3

**FIG. 2.7 :** Un exemple générique de chaîne de Markov à trois états représentée dans l'espace des états de $X$. Le modèle de transition représente les probabilités qui sont associées à chaque arc (non indiquées ici).



Or au vu de la simplification de la loi jointe de l'équation 2.3, il est également possible de représenter une chaîne de Markov dans le formalisme des réseaux bayésiens dynamiques comme indiqué sur la figure 2.8.

π X A X X X X

t=0 t=1 t=2 t=3 t=4

**FIG. 2.8 :** Un réseau bayésien dynamique associé à la chaîne de Markov de la figure 2.7.

Il est possible d'imaginer une chaîne de Markov d'ordre 2, c'est-à-dire pour laquelle l'état à l'instant $t$ dépendrait des états aux instants $t - 1$ et $t - 2$. Seulement, ce type de chaînes peut alors être vu comme une chaîne de Markov classique (d'ordre 1) ou nous considérons le vecteur aléatoire $Y_t = (X_{t-1}, X_t)$ comme variable aléatoire principale[ii] de cette chaîne.

biblio :

http://ofrancois.tuxfamily.org/Docs/these_Ofrancois.pdf

p