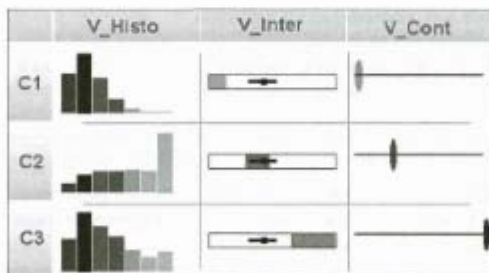


# Data Science par Analyse des Données Symboliques :

Une nouvelle façon d'analyser les  
données classiques, complexes et  
massives à partir des classes

Applications avec Syr et R



2018



Éditions TECHNIP 5 avenue de la République, 75011 PARIS

# Table des matières

<b>Avant-propos</b>	<b>V</b>
<b>Table des matières</b>	<b>1</b>
<b>Introduction générale</b>	<b>7</b>
<b>1. Des données classiques aux données symboliques</b>	<b>11</b>
1.1 Des individus aux classes	11
1.2 Des variables classiques aux variables symboliques	14
1.3 Processus d'agrégation et variables symboliques associées	22
1.4 Formalisation de quelques variables symboliques	27
1.4.1 Les variables multivaluées	28
1.4.2 Les variables à valeur intervalle	28
1.4.3 Les variables catégorielles multivaluées	29
1.4.4 Les variables modales	30
1.4.5 Les variables à valeur histogramme	30
1.4.6 Les variables à valeur diagramme de fréquences	31
1.5 Des variables classiques non appariées aux variables symboliques : le cas des données complexes	31
1.6 Intérêt de la description de classes par des données symboliques	33
1.7 Quelques principes de base	36
1.7.1 Classes considérées comme objets à décrire selon ses différentes facettes	36
1.7.2 La prise en compte de la variabilité interne aux classes	36
1.7.3 Dans l'interprétation, bien différencier les effets des causes	37
1.7.4 Le principe de généralisation	38
1.8 Quels sont les domaines d'application privilégiés de l'ADS ?	38
<b>2. Analyse descriptive pour des variables symboliques</b>	<b>41</b>
2.1 Statistiques élémentaires pour les variables à valeur intervalle	42

2.2	Statistiques élémentaires pour les variables à valeur histogramme.....	45
2.3	Calcul automatique d'histogrammes discriminants pour les classes.....	48
2.3.1	Discretisation pour une variable continue et algorithme de Fisher.....	50
2.3.2	Qu'est-ce que la discrimination entre classes ?.....	54
2.3.3	Une solution optimale : la méthode HistSyr.....	55
2.3.4	Une extension d'HistSyr au Big Data.....	57
<b>3.</b>	<b>Indices de proximité.....</b>	<b>75</b>
3.1	Dissimilarité pour une variable.....	76
3.1.1	Cas multivalué (Hausdorff, Minkowsky, Jaccard et Ichino).....	76
3.1.2	Cas modal ( $L^2$ et Hellinger).....	78
3.2	Dissimilarité entre plusieurs variables.....	79
3.3	La distance de Wasserstein.....	80
<b>4.</b>	<b>Classification automatique.....</b>	<b>83</b>
4.1	K-means et nuées dynamiques.....	84
4.2	La décomposition de mélange par partitions (MND) ou par partition floue (EM).....	87
4.2.1	Par la Méthode des Nuées Dynamiques (MND).....	87
4.2.2	Par l'algorithme d'Estimation-Maximisation (EM).....	87
4.2.3	Construction d'un tableau de données symboliques post Nuées dynamiques ou EM.....	89
4.3	Une extension de la méthode des nuées dynamiques aux données symboliques.....	92
4.3.1	Un choix de représentation d'une classe dans le cas de données symboliques.....	92
4.3.2	Critère d'affectation dans le cas symbolique où les noyaux sont des prototypes.....	93
4.3.3	Exemple.....	95
<b>5.</b>	<b>ACP étendue aux données symboliques.....</b>	<b>99</b>
5.1	Technique « classique ».....	100
5.2	Méthodes pour les variables à valeur intervalle.....	105
5.2.1	Technique par les centres.....	106
5.2.2	Technique par les sommets.....	107
5.2.3	Technique par les centres et les rayons.....	109

5.2.4	Technique par les fonctions de variance-covariance symboliques .....	113
5.3	Méthodes pour les variables à valeur histogramme .....	117
5.3.1	Technique par les variables « catégories », « globales » et « quadrants » .....	118
5.3.2	Technique par les moyennes après codage des catégories .....	126
<b>6.</b>	<b>Extension des règles d'association .....</b>	<b>133</b>
6.1	Règles d'association classiques et algorithmes d'extractions .....	134
6.1.1	L'algorithme Apriori pour l'extraction de règles d'association classiques .....	134
6.1.2	Extension des règles d'association classiques dans la littérature .....	139
6.2	Algorithme Apriori, règles d'association et données symboliques .....	143
6.2.1	Entrée de notre algorithme : un tableau de données symboliques .....	143
6.2.2	Objets symboliques et règles d'association symboliques .....	145
6.2.3	Définitions du support et de la confiance dans le cas de nos données symboliques .....	147
6.2.4	Algorithme Apriori symbolique (SApriori) .....	154
6.3	Règles d'association classiques versus symboliques .....	164
6.4	Complémentarité des règles d'association classiques et symboliques .....	167
<b>7.</b>	<b>Arbre de décision .....</b>	<b>175</b>
7.1	Description d'un arbre de décision classique .....	175
7.1.1	Les variables explicatives / à expliquer .....	177
7.1.2	Les nœuds terminaux / non terminaux .....	177
7.1.3	Ensembles d'apprentissage / de test .....	178
7.1.4	Entrées / sorties d'un arbre de décision .....	178
7.1.5	Construction et élagage d'un arbre .....	179
7.2	Extension des arbres de décision aux données symboliques .....	182
7.2.1	Les méthodes symboliques .....	182

7.2.2	Les arbres de décision étendus aux données symboliques avec la méthode SyrTree .....	185
7.2.3	Cas où la variable à expliquer est la classe (ou objet) symbolique .....	185
7.2.4	Cas où la variable à expliquer est une variable histogramme .....	191
7.3	Exemple illustratif .....	193
<b>8.</b>	<b>Prévision de données symboliques .....</b>	<b>199</b>
8.1	Régression linéaire à valeur intervalle .....	199
8.1.1	Méthode classique .....	200
8.1.2	Méthode par les centres des intervalles .....	202
8.1.3	Méthodes par les centres et les étendues des intervalles : cadre univarié et bivarié .....	203
8.1.4	Méthode par les fonctions de variance-covariance symboliques .....	206
8.1.5	Comparaison des méthodes .....	208
8.2	Régressions linéaires pénalisées à valeur intervalle .....	209
8.2.1	Régressions « ridge », « lasso » et « elastic standard » .....	209
8.2.2	Régressions pénalisées pour les intervalles .....	211
8.3	Séries temporelles à valeur intervalle .....	215
8.3.1	Cas classique et stationnaire (linéaire et non linéaire) : rappels .....	215
8.3.2	Méthodes symboliques : cadre stationnaire .....	226
8.3.3	Méthode symbolique par les k plus proches voisins : cas non stationnaire .....	229
8.3.4	Comparaison des méthodes .....	230
<b>9.</b>	<b>Exercices d'application avec R et Syr .....</b>	<b>231</b>
9.1	Variabilité externe pour des données en finance de type intervalle .....	231
9.2	Statistiques descriptives sur des données en finance de type intervalle .....	235
9.3	ACP sur les cours « Lehman Brothers » de type intervalle .....	240
9.4	ACP sur les cours « Lehman Brothers » de type histogramme .....	252
9.5	Distance de Wasserstein sur les données de prismes .....	257

9.6	Régression sur intervalles pour les cours « Lehman Brothers » .....	263
9.7	Prévision journalière et par intervalle des cours « Lehman Brothers » .....	277
<b>10.</b>	<b>Applications avec Syr et R .....</b>	<b>293</b>
10.1	Étude de la dégradation des tours de refroidissement des centrales nucléaires EDF .....	293
10.1.1	Les mesures de surveillance de la dégradation des tours de refroidissement .....	294
10.1.2	Étude de la dégradation des tours et comparaison des tours entre elles .....	296
10.1.3	Conclusion .....	305
10.2	Étude de l'influence des conditions environnementales sur la corrosion du béton armé .....	306
10.2.1	Programme expérimental .....	306
10.2.2	Données symboliques et nouveaux seuils de corrosion .....	309
10.2.3	Proximités entre agressions : l'exemple de la variable Ecorr .....	316
10.2.4	Résultats sur les corrélations entre Icorr et Ecorr, et entre Icorr et Re .....	319
10.3	Classification et arbre de décision pour les trajectoires de prise en charge des patients atteints d'un cancer du colo-rectum .....	321
10.3.1	Données .....	322
10.3.2	Résultats .....	325
10.3.3	Bilan .....	328
10.3.4	Conclusion .....	329
10.4	Extraction de thématiques sur un corpus de documents issus d'appels téléphoniques .....	331
10.4.1	Présentation des données initiales .....	331
10.4.2	Objectifs de l'étude et stratégie de résolution .....	332
10.4.3	Construction, visualisation et classification des tableaux de données .....	334
10.4.4	Sélection automatique des mots d'intérêt .....	337
10.4.5	Conclusion .....	339



10.5 « Symbolic covariance ACP » et régression sur des données de type intervalle en épidémiologie vétérinaire .	340
10.5.1 Données .....	340
10.5.2 Statistiques pour variables à valeur intervalle .....	342
10.5.3 Résultats de la « Symbolic Covariance PCA » .....	343
10.5.4 Résultats de la « Symbolic Covariance Regression » .....	346
10.5.5 Conclusion.....	347
10.6 Mesures de Value at Risk à valeur histogramme : une approche symbolique pour l'attribution du risque .....	348
10.6.1 À propos des fonds alternatifs et des mesures de VaR.....	348
10.6.2 Traitements sur les données .....	349
10.6.3 Résultats de l'ACP par les variables « globales » et « catégories » .....	352
10.6.4 Résultats de la classification par nuées dynamiques.....	355
10.6.5 Conclusion.....	360
10.7 Analyse des données de capteurs (Big Data).....	361
10.7.1 Présentation des données initiales .....	361
10.7.2 Recherche des histogrammes les plus discriminants avec CloudHistSyr.....	362
<b>Conclusion .....</b>	<b>373</b>
<b>Annexe 1 : le logiciel SYR.....</b>	<b>375</b>
<b>Annexe 2 : des modules de R pour l'ADS.....</b>	<b>385</b>
<b>Annexe 3 : des pistes de recherche et de développement ...</b>	<b>409</b>
<b>Bibliographie .....</b>	<b>419</b>