Exploration et analyses multi-objectifs de séries temporelles de données météorologiques

Yelen PER*, Kevin DALLEAU**, Malika SMAIL-TABBONE***

*LORIA UMR 7503, CNRS, yelen.per@loria.fr
**LORIA UMR 7503, CNRS, kevin.dalleau@loria.fr
***LORIA UMR 7503, Université de Lorraine, malika.smail@loria.fr

Résumé. Cet article présente les investigations menées sur les données mesurées par des capteurs positionnés dans cinq villes de l'île de la Réunion. Des analyses exploratoires préalables permettent de comparer les caractéristiques statistiques des villes considérées relativement aux différentes variables météorologiques mesurées (flux solaires diffus et global, pression atmosphérique, humidité, température, force et direction du vent). Nous appliquons diverses transformations sur les données avant d'analyser les séries univariées ou multivariées agrégées au pas de l'heure ou de la journée afin de construire des modèles de prédiction. Une approche classique de clustering de séries temporelles est testée. Deux algorithmes de biclustering appliqués successivement ont permis de grouper les journées d'observations partageant des paramètres météorologiques horaires. Une caractérisation des biclusters, une visualisation calendaire de leur succession ainsi qu'une recherche de séquences fréquentes permettent d'exploiter les résultats et de faciliter leur interprétation.

1 Introduction

Dans le cadre du défi EGC 2018, deux années de mesures - au pas de la minute - de variables météorologiques dans cinq villes de l'île de la Réunion ont été mises à disposition en vue de leur analyse. Le projet à l'origine de ces données s'inscrit dans le cadre de la politique de développement vers l'autonomie énergétique de l'île. Nous avons choisi d'une part de construire, sur la base des séries de données multivariées, des modèles de prédiction ou *prévision* de l'indice de fraction directe et d'autre part, d'appliquer des méthodes de classification non-supervisée (clustering) sur des données univariées puis multivariées décrivant les journées. Un outil de visualisation permet de voir comment des groupes de journées à profil météorologique similaire se succèdent sur un calendrier.

Les éléments de l'analyse exploratoire des données sont présentés avec les principales transformations des données brutes dans la section 2. Les méthodes de régression utilisées et quelques résultats obtenus sont présentés dans la section 3. Les sections 4 et 5 exposent les deux expériences de clustering ainsi que l'outil de visualisation réalisé pour l'occasion.

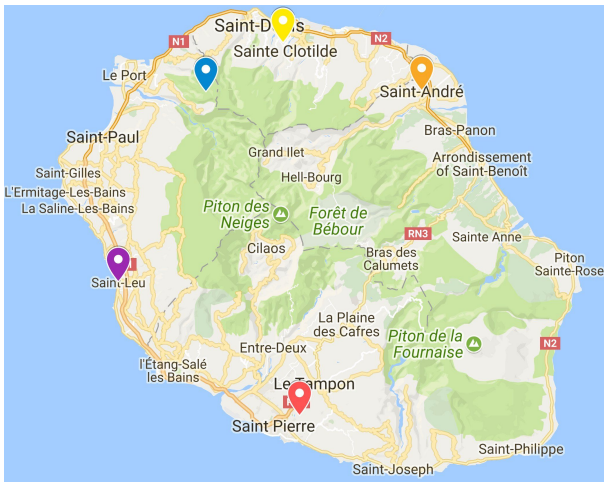


FIG. 1 – *Positionnement*[1] *des villes d'intérêt : La Possession (P) (bleu), Saint-Leu (SL) (violet), Saint-Pierre (SP) (rouge), Saint-André (SA) (orange) et Moufia (M) (jaune).*

2 Analyse exploratoire des données et principales transformations

Sept variables météorologiques ont été mesurées en 2014 et 2015 au pas de la minute dans cinq stations de la Réunion : température extérieure (Text), pression atmosphérique (Patm), taux d'humidité (RH), force et direction du vent (WS, WD), flux solaires global et diffus (FG, FD). L'indice de fraction directe k_b est une variable importante dans le contexte météorologique et est défini comme le rapport entre le flux solaire direct (lui-même défini comme la différence entre flux global et flux diffus) et le flux global. Nous avons adjoint aux sept variables l'indice k_b. Cet indice, compris entre 0 et 1, est proportionnel au degré d'ensoleillement. Ainsi un indice proche de 0 indique une journée nuageuse tandis qu'un indice proche de 1 indique une journée ensoleillée (et donc susceptible de fournir une grande quantité d'énergie).

Les stations concernées par les données sont Moufia (au nord), La Possession (au nord-ouest), Saint-André (à l'est), Saint-Leu (à l'ouest) et Saint-Pierre (au sud) (cf. figure 1).

Quelques valeurs très extrêmes - notamment des flux diffus et global mesurés en janvier 2014 - rendent difficile la visualisation des statistiques. Nous avons également noté que les valeurs de k_b ne sont pas cantonnées dans l'intervalle $[0, 1]$. Cela est probablement dû à l'incertitude des capteurs de flux solaires utilisés qui s'élève à 11% pour la mesure du flux diffus et autant pour celle du flux global, ce qui nous amène à une incertitude de plus de 20% pour la valeur de l'indice k_b. La figure 2 montre les statistiques de base (minimum, maximum, médiane, premier et troisième quartile, moyenne marquée par un point vert) après suppression des valeurs très extrêmes pour chaque année (2014 en rouge et 2015 en bleu).

1. Carte réalisée avec Google Maps.

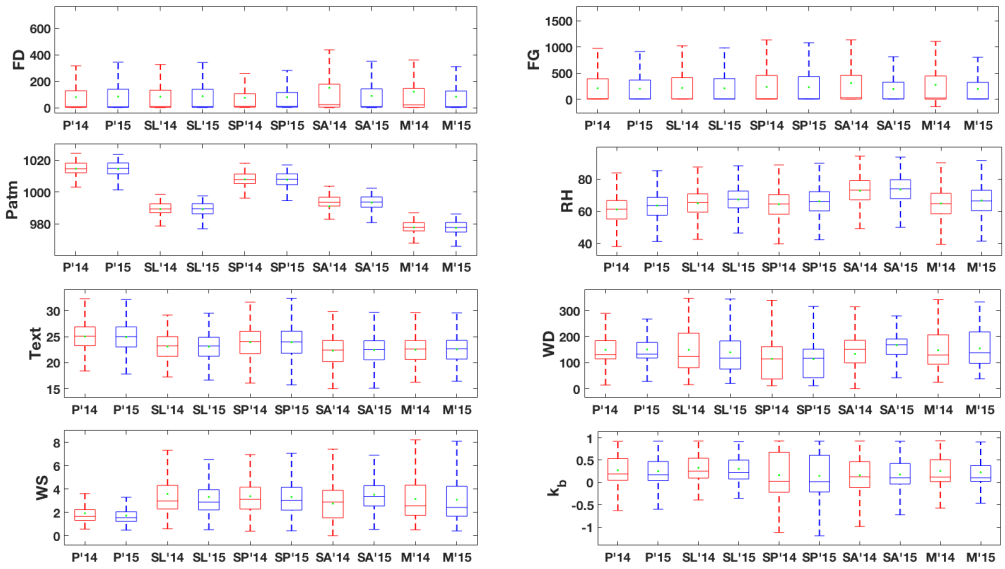


FIG. 2 – *Statistiques sur les données brutes dans les cinq stations après suppression des valeurs extrêmes. Les noms des stations ont été abrégés : M'14 correspond à Moufia 2014.*

Ces statistiques montrent quelques différences entre les cinq villes telles qu'une température légèrement supérieure et un vent moins fort à La Possession, un taux d'humidité supérieur à Saint-André, une pression atmosphérique plus élevée à La Possession et à Saint-Pierre. La période de deux ans est évidemment trop courte pour identifier des tendances dans les séries de mesures, hormis une légère augmentation de la force du vent à Saint-André.

La figure 3 montre les indices de corrélation entre les différentes variables (où les flux solaires global et diffus ont été omis au profit de l'indice k_b) pour chacune des cinq villes. On constate ici que les corrélations entre variables sont similaires pour les villes de La Possession, de Saint-Leu et de Saint-Pierre. Cela correspond bien à la situation géographique de ces villes, situées sur la côte sous le vent [2]. Les villes situées sur la côte au vent se démarquent par une corrélation positive entre la direction du vent et k_b. Saint-André se démarque de toutes les villes par une corrélation négative entre pression atmosphérique d'une part et température extérieure et taux d'humidité d'autre part.

Des jeux de données ont été constitués pour chaque ville par transformation des données brutes au pas de la minute ($\sim$ 1 044 000 lignes). Une agrégation horaire est réalisée et nous générons un jeu de données multivariées au pas de l'heure dans lequel les mesures de chaque heure sont moyennées ($\sim$ 17 400 lignes). Une agrégation journalière donne un jeu de données univariées par variable météorologique où une journée est représentée par la moyenne de la variable considérée (pression atmosphérique, k_b...) sur les heures de la journée. Chaque jeu de données journalières univariées comporte 729 lignes et 24 colonnes ou 13 colonnes si l'on se limite aux mesures diurnes (de 6h à 19h).

2. Le climat de La Réunion est – entre autres – marqué par une différence entre deux côtes, la *côte au vent* et la *côte sous le vent*. Cette dernière est protégée des alizés, vents dominants, par des montagnes.

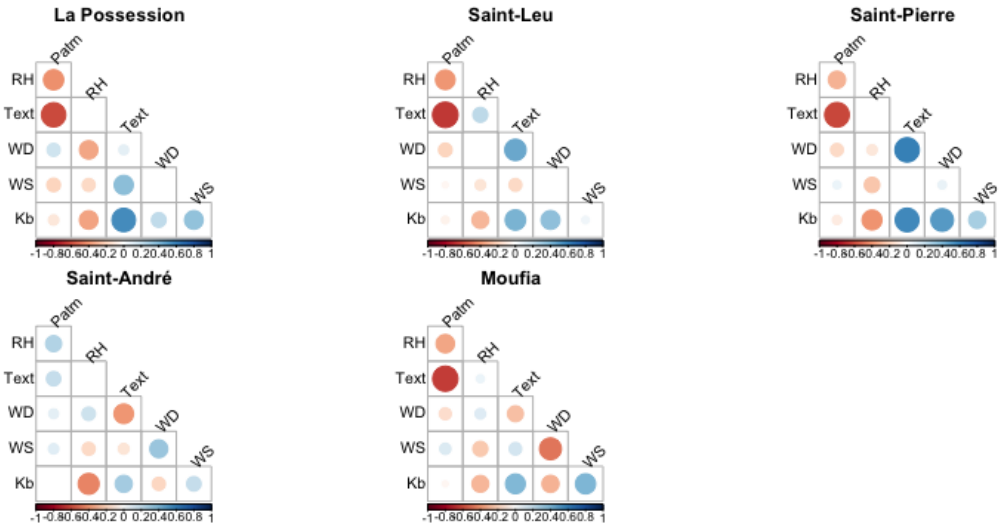


FIG. 3 – *Corrélations entre variables pour les villes étudiées.*

3 Prévision de l'indice de fraction directe à partir de données multivariées

Afin de permettre la prévision de l'indice de fraction directe à partir de mesures passées, nous avons choisi de travailler à l'échelle de l'heure [3]. Nous avons ensuite enrichi les données horaires multivariées de chaque ville en ajoutant des décalages temporels pour chaque variable météorologique. À ce stade, les lignes du jeu de données avec une valeur de k_b en dehors de l'intervalle $[0, 1]$ ont été supprimées, ce qui représente 20 à 25% selon le jeu de données (ou selon la ville). Dès lors, il s'agit simplement de prédire, par une méthode de régression, la valeur de k_b au temps H à partir des valeurs de k_b, FD, FG, Patm, Text, RH, WD et WS aux temps H-1, H-2, H-3 et H-4 (des décalages plus grands n'ont pas montré d'amélioration des résultats).

Nous avons construit et comparé les performances de trois modèles de régression implémentés dans la plateforme Weka (Witten et Frank (2005)). Le modèle linéaire (*SimpleLinearRegression*) a servi de base de comparaison avec un arbre de décision (*RepTree*) (Holmes et al. (1999)) et un ensemble de règles issues d'arbres de modèles linéaires (*M5Rules*) (Quinlan (1992)).

Un test statistique de Student apparié a été appliqué pour comparer les valeurs de deux métriques sur 100 expériences (10 répétitions de validations croisées à 10 plis) : *(i)* le coefficient de corrélation statistique entre valeurs mesurées de k_b et valeurs prédites et *(ii)* la racine carrée de la moyenne du carré de l'erreur (RMSE). Les résultats des tests effectués sur les données de La Possession et Moufia sont présentés en table 1. Ils sont en faveur du modèle linéaire pour RMSE mais en faveur de l'arbre de décision et de l'ensemble des règles quand on

3. Des tests non satisfaisants ont été effectués à l'échelle de la journée. Cela est probablement dû à la faible taille du jeu de données et à l'imprécision des valeurs agrégées sur la journée.

	SimpleLinearRegression		*RepTree*		*M5Rules*	
	Coefficient de corrélation	RMSE	Coefficient de corrélation	RMSE	Coefficient de corrélation	RMSE
Données Moufia	0.81	0.17	**0.87**	0.13	**0.88**	0.14
Données La Possession	0.76	0.18	**0.86**	0.15	**0.87**	0.14

TAB. 1 – *Résultats obtenus sur 10 répétitions de validations croisées à 10 plis de trois programmes de régression (avec les paramètres par défaut) sur deux jeux de données. Les valeurs en gras indiquent le programme dont les résultats sont significativement meilleurs que la baseline selon la métrique choisie pour le test.*

effectue le test par rapport au coefficient de corrélation. Les valeurs moyennes de RMSE (entre 0.13 et 0.18) sont acceptables compte tenu de l'incertitude de 20% qui entâche la mesure de k_b. Les valeurs des coefficients de corrélation sont très bonnes (au moins 80%), indiquant qu'il est possible de proposer une prévision fiable des variations de l'indice de fraction directe sur la base des données telles que nous les avons mises en forme.

4 Clustering des journées sur la base de données univariées

En plus de la problématique de prévision, il est pertinent de s'intéresser à l'identification automatique de journées types par rapport à l'ensoleillemment afin de tenter de cerner les spécificités des différentes villes de la Réunion. Pour ce faire, nous nous focalisons sur des séries univariées, au pas de l'heure, par ville. La variable considérée est k_b. Les données nocturnes ont été censurées, afin de ne garder que les données intéressantes concernant le flux solaire. La mesure de distance entre séries choisie est la mesure DTW (*dynamic time warping*). L'objectif est de grouper les journées présentant le même profil d'évolution de l'indice de fraction directe au sein d'un même cluster. Certaines villes, notamment Moufia et Saint-André, présentent des mesures extrêmes, pouvant perturber le clustering. Nous avons donc opté pour une méthode de clustering résistante aux *outliers*, à savoir l'algorithme *genie* (Gagolewski et al. (2016)). La réalisation d'un clustering avec cette méthode nous permet d'obtenir un dendrogramme pour chaque ville. L'examen de ces dendrogrammes nous a permis de faire quelques constats :

— Les villes de La Possession, de Saint-Leu et de Saint-Pierre, toutes trois villes de la côte sous le vent, ont des dendrogrammes similaires, avec des partitions bien marquées pour des hauteurs d'agglomération h aux alentours de 1.7.
— Une coupure très nette autour de $h = 2.25$ pour Moufia.
— Un dendrogramme bien spécifique, avec des clusters très homogènes pour la ville de Saint-André.

Pour des raisons d'espace, nous ne présentons que les dendrogrammes de Saint-André et de Saint-Pierre (figures 4 et 5).

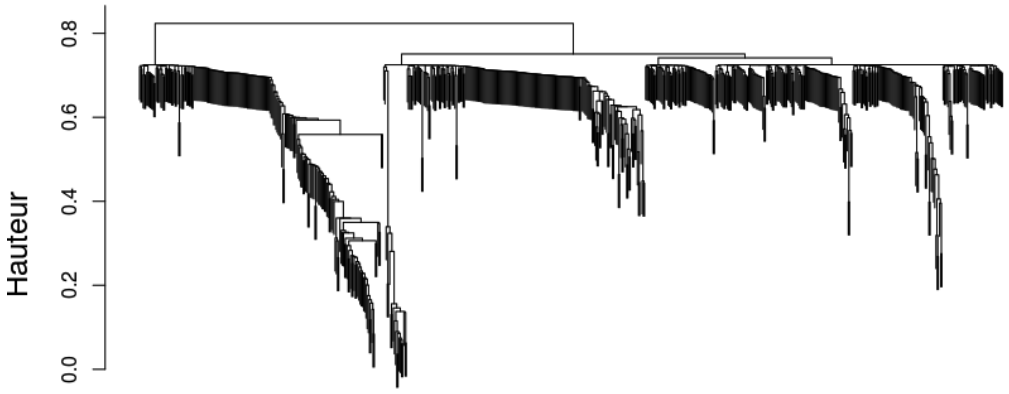


FIG. 4 – *Dendrogramme correspondant aux k_b pour la ville de Saint-André.*

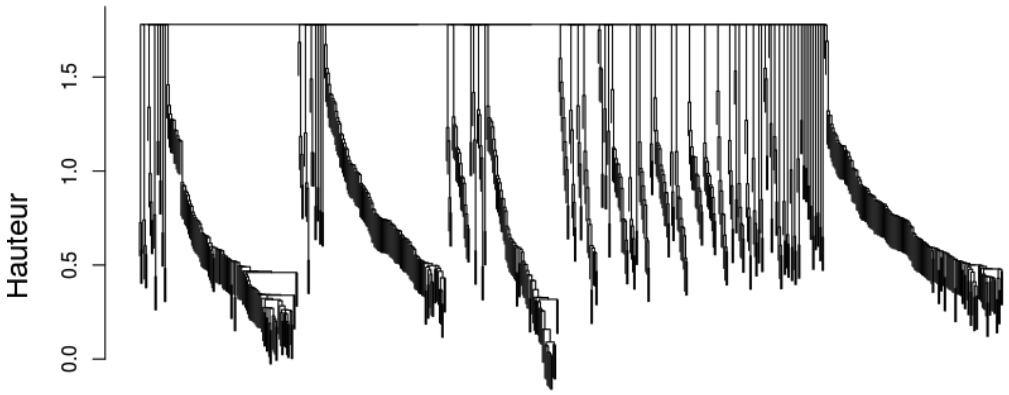


FIG. 5 – *Dendrogramme correspondant aux k_b pour la ville de Saint-Pierre.*

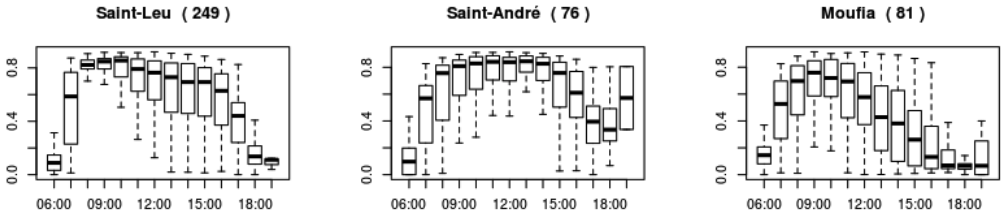


FIG. 6 – *Évolution de k_b pour les journées des clusters les plus gros dans trois villes. Le nombre de jours concernés est indiqué entre parenthèses.*

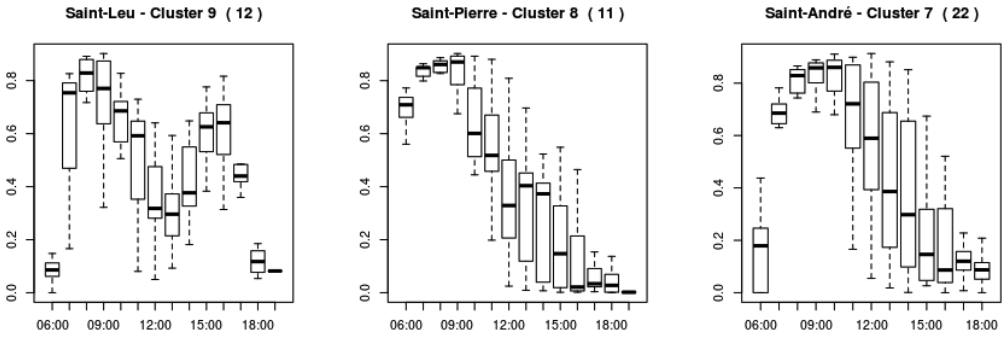


FIG. 7 – *Évolution de k_b pour les journées de certains clusters spécifiques dans trois villes. Le nombre de jours concernés est indiqué entre parenthèses.*

Un clustering est défini sur les villes de la côte sous le vent, en coupant les dendrogrammes à $h = 1.7$. Cette coupure donne 67 clusters pour La Possession, 68 clusters pour Saint-Leu et 59 clusters pour Saint-Pierre. Le nombre de clusters obtenus est élevé, et semble suggérer que l'évolution de k_b au cours des journées est très variable dans chacune de ces villes. De nombreux clusters ne contenant que quelques jours, nous avons fait le choix de ne considérer que les clusters contenant plus de 10 jours. Le même clustering est réalisé pour les villes du Moufia ($h = 0.62$) et de Saint-André ($h = 0.725$). Quatre et neuf clusters sont obtenus, respectivement. La figure 6 présente l'évolution moyenne des valeurs de k_b dans les clusters contenant le plus d'éléments pour Saint-Leu, Saint-André et Moufia. Nous avons fait le choix de ne pas présenter les courbes correspondant à La Possession et à Saint-Pierre car très proches de celle de Saint-Leu. Ces clusters représentent les journées types pour chacune des villes concernées. La journée type ne diffère donc que très peu d'une ville de l'île à une autre. Nous nous sommes donc orientés vers la comparaison de clusters apparaissant moins fréquemment. Certains de ces clusters sont intéressants car associés à des journées très ensoleillées comme on peut le voir sur la figure 7.

Bien que ces clusters spécifiques permettent de décrire certains types de journées, ceux-ci sont obtenus sur la base d'une seule variable. Une étude multivariée a donc été réalisée pour le clustering des journées.

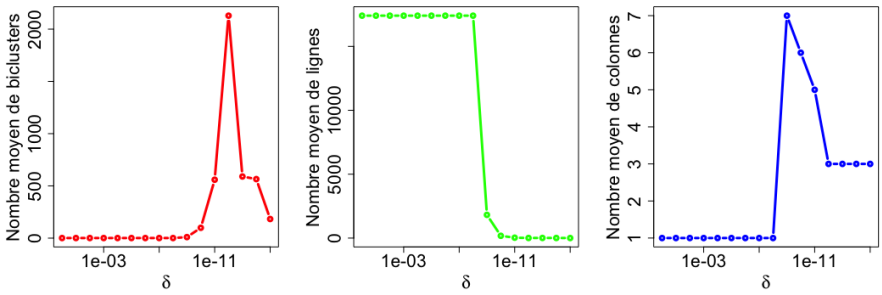


FIG. 8 – *Évolution, en fonction de différentes valeurs du seuil δ, des nombres moyens de biclusters, de lignes et de colonnes par bicluster.*

5 Clustering des journées sur la base de données multivariées

Nous sommes ici intéressés par la définition de profils météorologiques à l'échelle de la journée. Afin de prendre en compte les données multivariées nous avons eu recours à des méthodes de biclustering (ou co-clustering) qui ont l'avantage d'être moins rigides en n'imposant pas une similarité sur la globalité des colonnes pour regrouper des objets (lignes).

5.1 Étape 1 : biclustering des heures sur la base des données météorologiques numériques

Des groupes d'heures associées à des amplitudes similaires des variables météorologiques (FD, FG, Text, Patm...) ont été recherchés avec l'algorithme de biclustering de Cheng & Church (Cheng et Church (2000)). Celui-ci considère initialement une matrice $M = I$ de données et optimise, relativement à une valeur seuil δ, un score H basé sur une adaptation de l'erreur quadratique moyenne associée à une sous-matrice de données ou bicluster. D'une part, les lignes et colonnes de I contribuant le plus à H sont supprimées de M. D'autre part, les lignes et colonnes de I contribuant le moins à H sont ajoutées à M. Après remplacement des éléments de M dans I par des valeurs aléatoires, la réapplication de l'algorithme sur I permet de rechercher une nouvelle sous-matrice. Les biclusters finaux ne couvrent pas forcément l'intégralité des données et peuvent se chevaucher.

Le score d'erreur optimisé est calculé à partir des données. Afin d'éviter une influence inégale des facteurs climatiques, les données ont préalablement été normalisées entre 0 et 1. Un seuil δ optimal a ensuite été recherché par expérimentation. Les nombres moyens par ville de biclusters, de lignes par bicluster et de colonnes par bicluster ont été relevés (cf. figure 8). Le nombre moyen de colonnes par bicluster est maximal pour $\delta = 10^{-9}$, valeur pour laquelle le nombre moyen de biclusters est 10, le nombre moyen de lignes par bicluster vaut 1 817 et le nombre de colonnes est de 7. Les données contenant environ 17 400 lignes, un bicluster couvre ainsi environ 10% des données, ce qui en fait un ensemble assez significatif. Nous obtenons donc 10 biclusters d'environ 1 800 heures partageant un profil similaire pour l'ensemble des paramètres météorologiques.

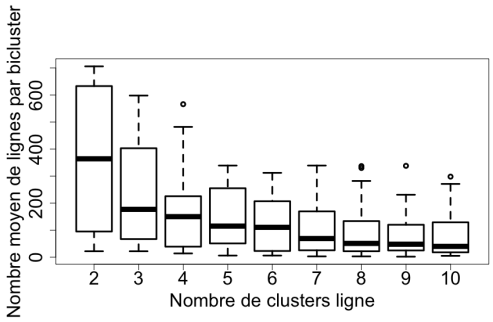


FIG. 9 – *Évolution, en fonction du nombre de clusters ligne, du nombre moyen de lignes par bicluster.*

5.2 Étape 2 : biclustering symbolique des données journalières

Une fois les biclusters d'heures obtenus, nous les utilisons pour représenter chaque journée par la succession des biclusters attachés à chaque heure de la journée (de 0 à 23). Analyser ces données symboliques afin de grouper les journées ayant un profil météorologique similaire (vecteur de biclusters) peut être vu comme un problème de biclustering avec la contrainte de garder toutes les colonnes dans les biclusters afin de préserver l'intégrité de la série temporelle de chaque journée.

L'algorithme du modèle à blocs latents (Keribin et al. (2015)) a été utilisé, notamment parce qu'il permet de fixer un nombre de clusters ligne (nombre de sous-ensembles de lignes) et un nombre de clusters colonne (nombre de sous-ensembles de colonnes) à trouver dans une matrice. Par ailleurs les clusters obtenus couvrent l'intégralité des données et ne se chevauchent pas.

Dans notre cas, le nombre de clusters colonne est fixé à 1 afin de prendre en compte l'ensemble des heures de la journée. Le nombre moyen de lignes par bicluster est analysé pour fixer le nombre de clusters ligne (cf. figure 9). L'algorithme paramétré étant stochastique, afin d'assurer un résultat stable au fil des exécutions, le nombre de six clusters ligne semble être un bon compromis entre nombre de clusters, tailles des clusters et amplitude de ces tailles.

5.3 Analyse et visualisation des biclusters obtenus

Une représentation visuelle et compacte de la répartition calendaire des biclusters de journées est proposée (figure 10). Chacun des biclusters peut être caractérisé par les valeurs médianes des différents facteurs climatiques. Le nombre d'apparition sur une année ou un mois de chacun des biclusters est par ailleurs calculé.

Afin de capter des relations entre biclusters, ou des transitions climatiques récurrentes, des séquences fréquentes de biclusters sont intéressantes à rechercher. L'algorithme *PrefixSpan* (Pei et al. (2001)) procédant par projection de motifs sur des items fréquents a été utilisé. Une fenêtre de taille W est glissée sur la séquence des biclusters et des sous-séquences fréquentes sont recherchées. Des tests nous ont permis de choisir $W = 5$ de façon à obtenir des motifs à support à la fois élevé et stable.

Un programme de visualisation [4] a été développé et peut être testé en ligne. Ce programme requiert le chargement d'un jeu de données (les jeux de données mis à disposition sont préchargés sur le serveur), y applique les traitements décrits (la figure 11 résume ces traitements) et affiche la séquence calendaire, les caractéristiques médianes, les nombres d'apparition ainsi que les sous-séquences fréquentes des biclusters identifiés. L'utilisation de l'outil a permis de mettre en exergue quelques faits :

— Moufia et La Possession sont sujets à des cyclones (ces phénomènes entraînent une couverture nuageuse), notamment Bejisa les 2 et 3 janvier 2014 et Haliba vers le 9 mars 2015.
— Moufia et Saint-André ont connu une période de sécheresse (de telles périodes sont caractérisées par des flux solaires importants, un flux global environ 1.5 fois supérieur au flux diffus et une température élevée) début janvier 2014.
— Saint-André et Saint-Leu connaissent des saisons relativement bien marquées avec des intersaisons en avril et en septembre.
— Le sud-ouest de l'île semble bénéficier d'un climat plus doux que le nord-est.

L'analyse des séquences fréquentes de biclusters a abouti aux résultats suivants :

— Des journées ensoleillées, couvertes par moment, caractérisent le climat de La Possesion, positionnée au nord-ouest de l'île.
— Des transitions climatiques marquées sont typiques des villes de Saint-Leu et Saint-Pierre, localisées dans le sud de l'île.
— Des journées couvertes, ensoleillées par moment, sont représentatives des villes de Saint-André et Moufia, situées au nord de l'île.

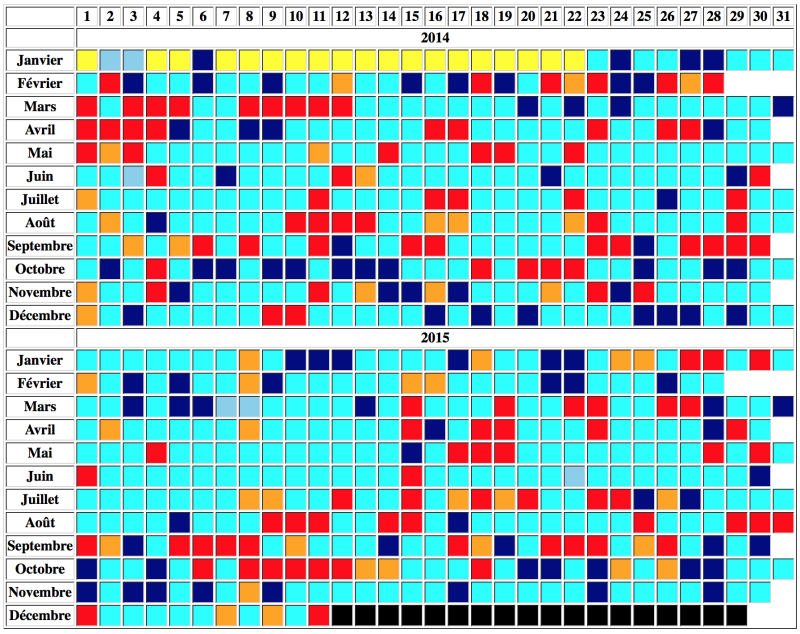
6 Conclusion et perspectives

Nous avons mené de nombreuses expériences sur les données du défi de la conférence EGC 2018. Nous présentons dans cet article quelques éléments d'analyse et quelques résultats de façon à rendre compte de la richesse de ces données et des questions qu'elles suscitent (y compris auprès de néophytes). La multiplicité des jeux de données, des variables et des possibilités d'agrégation conduit à un espace assez conséquent d'analyses exploratoires possibles. Une interaction avec des experts des données nous aurait certainement aidés à mieux interpréter les résultats obtenus et probablement orientés vers d'autres expériences. Par exemple, un autre scénario de régression potentiellement intéressant serait de tenter de prédire l'indice de fraction directe d'une journée à partir des valeurs météorologiques mesurées pendant les premières heures du jour.

Par ailleurs, une discrétisation experte des variables météorologiques journalières/horaires aurait permis une recherche de règles d'associations susceptibles de refléter des interactions plus complexes (que des coefficients de corrélation) entre variables météorologiques pour des ensembles significatifs de jours/heures.

Nous avons développé un outil de visualisation que nous avons doté de diverses fonctionnalités destinées à tirer profit des résultats des analyses. Une version étendue (par exemple à d'autres façons de définir les biclusters) pourrait servir de tableau de bord pour aider à la décision.

4. https://avicenne-test1.loria.fr



	Indice de fraction directe	Flux diffus (W/m²)	Flux global (W/m²)	Pression atmosphérique (hPa)	Humidité (%)	Température (°C)	Direction du vent (°)	Force du vent (m/s)
	0.12	48	82	975	71	25	177	1
	0.15	6	6	983	65	21	98	6
	0.26	31	73	977	61	25	93	3
	0.31	175	243	976	73	24	126	2
	0.42	30	44	975	60	25	94	4
	0.47	31	39	977	66	25	95	4

FIG. 10 – *Répartition calendaire et caractéristiques médianes de biclusters pour la ville du Moufia. Les cases blanches indiquent une absence de données. Les cases noires indiquent une absence de bicluster.*

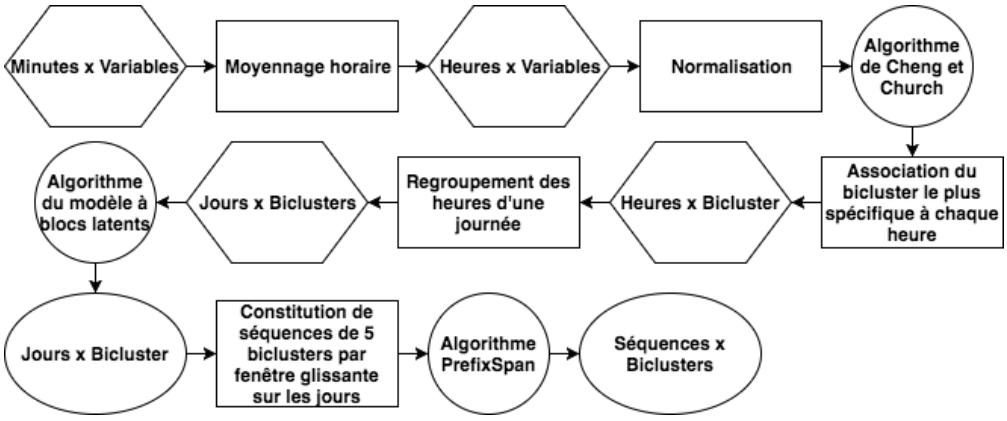


FIG. 11 – *Données (hexagones), transformations (rectangles), algorithmes (cercles) intervenant dans le traitement et résultats (ellipses).*

Références

Cheng, Y. et G. M. Church (2000). Biclustering of expression data. In P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, et H. Weissig (Eds.), *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, August 19-23, 2000, La Jolla / San Diego, CA, USA*, pp. 93–103. AAAI.

Gagolewski, M., M. Bartoszuk, et A. Cena (2016). Genie : A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences 363*, 8–23.

Holmes, G., M. Hall, et E. Frank (1999). Generating rule sets from model trees. In *In Proc. 12th Australian Joint Conference on Artificial Intelligence (AI-99*, pp. 1–12. Springer.

Keribin, C., V. Brault, G. Celeux, et G. Govaert (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing 25*(6), 1201–1216.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefixspan : Mining sequential patterns by prefix-projected growth. In D. Georgakopoulos et A. Buchmann (Eds.), *Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany*, pp. 215–224. IEEE Computer Society.

Quinlan, R. J. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343–348. World Scientific.

Witten, I. et E. Frank (2005). *Data Mining : Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.

Summary

Two years of per minute data, measured by sensors located in five cities of Reunion Island, are investigated in this paper. Prior exploratory data analyses have enabled the statistical comparison of characteristics of cities with respect to the measured weather variables (diffuse and overall solar fluxes, atmospheric pressure, moisture, temperature, wind speed and direction). Data was preprocessed and univariate time-series and multivariate time-series aggregated over hours or days were analyzed in order to build simple and effective prediction models. A classical clustering approach was performed. Groups of days sharing weather parameters in common were found by two biclustering algorithms. The characterisation of found biclusters and their succession displayed in a calendar-based visualization tool have helped assess their interest.