Reconstitution des séries mensuelles de températures maximales et minimales sur l'ouest Algérien

Karima SOLTANI [1]*,Mahmoud HAOUARI [1]

Résumé

[1] *Département Météorologique Régional Ouest (ONM, Algérie)*
***Correspondant**: kari.mekki@yahoo.com

1. Introduction

Toutes les études relatives au changement climatique et à la variabilité climatique nécessitent de longues séries climatologiques d'une très bonne qualité. Dans notre pays, malgré que les premières observations météorologiques aient débutées vers 1850 au port d'Alger, il n'en demeure pas que ces séries présentent beaucoup de lacunes et d'incohérences. Il est temps pour l'ONM de se pencher sérieusement sur cet aspect de sa banque de données pour entamer la lourde tâche de construire les plus longues séries des paramètres climatologiques importants (température, précipitation, pression atmosphérique ...).

Aujourd'hui, le software existe en open source sous R Team (2014) pour effectuer l'ensemble de ces tâches fastidieuses, qui nous étaient inaccessibles sans l'aide de pays très avancés dans ce domaine. Les meilleurs programmes d'imputation des données manquantes et d'homogénéisation des séries climatologiques existent par dizaines en libre utilisation.

C'est dans ce cadre, que s'inscrit ce modeste travail pour montrer à nos collègues de l'ONM, les possibilités qui s'offrent maintenant pour nous pour relever certains défis, que nous seuls peuvent relever pour la sauvegarde du patrimoine climatique de notre pays.

2. Objectif

Jusqu'à présent dans tout l'ouest Algérien, l'ONM ne dispose que d'une seule série d'Oran Es-sénia, de température mensuelle de 1936 à nos jours qui est complète et de bonne qualité. Toutes les autres séries sont lacunaires. Ces lacunes, comme le montre le tableau 1, sont imputées à la période post coloniale de 1962 jusqu'à la fin des années 70. Après cette période, le réseau professionnel s'est largement étoffé pour reprendre l'exploitation des anciennes stations et d'en ouvrir d'autres.

Il s'avère que la seule série disponible d'Oran Es-Sénia est très insuffisante pour caractériser l'évolution des températures dans l'ouest Algérien, car elle n'est représentative que du climat du littoral.

Depuis déjà, un certain temps que nous réfléchissons à constituer d'autres séries assez complètes pour enrichir nos connaissances, mais malheureusement, le rattrapage des données manquantes était impossible avec une seule station.

Récemment l'idée nous ai venu d'utiliser les champs analysés des températures de surface du projet ERA40 du centre météorologique européen ECMWF Uppala et al. (2005). Ces champs s'échelonnent de 1958 à 2002, ce qui nous convient parfaitement parce qu'ils couvrent notre période lacunaire.

Enfin, à partir de ces hypothèses, l'idée de constituer des séries mensuelles de températures maximales et minimales du tableau 1, était devenu plus claire.

Table 1. Fonctionnement des stations utilisées dans l'étude

Station	Période des séries	Arrêts de fonctionnement
Oran (Es-sénia)	1936-2013	aucun
Mascara (Ghriss)	1958-2013	09/1959 à 06/1983
Sidi bel abbes	1957-2013	04/1964 à 12/1984
Tlemcen Zenata	1958-2013	11/1962 à 09/1980
El kheiter	1942-2013	11/1962 à 02/1978
Mecheria	1958-2013	11/1962 à 10/1979
El bayadh	1948-2013	07/1962 à 09/1971

3. Matériel et méthodes

Exposé de la méthode théorique d'imputation

Les méthodes d'imputation des données manquantes se divisent généralement en trois groupes

- Méthodes d'imputation simple
- Méthodes d'imputation multiple
- Méthodes basées sur le maximum de vraisemblance

Les méthodes basées sur le maximum de vraisemblance (likelihood-based methods) peuvent cependant être aussi bien utilisées en simple qu'en multiple imputation.

Dans ce travail, nous utiliserons une méthode d'imputation simple par Analyse en Composante Principale.

On peut définir l'ACP comme une projection orthogonale des données dans un sous espace linéaire (espace principal) dans lequel la variance des données est maximisée, c'est-à-dire qui préserve le maximum d'informations.

On peut montrer que la projection optimale dans le sous espace de k dimensions ou l'on a choisit k vecteurs propres $\{w_j\}, j = 1,...,k$ de la matrice de covariance S correspondent aux k plus grandes valeurs de valeurs propres $(\lambda_1,...,\lambda_k)$.

La matrice de covariance est donnée par

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})(x_n - \overline{x})^T \qquad (1)$$

Maintenant, la transformation linéaire de la matrice des données x_n dans le sous-espace principal définit par les k vecteurs propres est simplement :

$$z_n = W^T(s_n - \mu) \quad (2)$$

Où z_n est la matrice des composantes principales, W^T est la transposée de la matrice formée par les k vecteurs propres et μ est le terme biais qui représente la moyenne des données.

Une autre propriété de l'ACP est qu'elle donne une représentation linéaire des données en k dimensions , de telle sorte que l'erreur quadratique des données reconstruites par $\widehat{x}_n = Wz_n + \mu$ est minimale.

Enfin, l'analyse en composante principale probabiliste(PPCA) Vatanen (2012) permet de contourner la limitation de l'analyse en composante principale classique en introduisant le notion de distribution de probabilité en ajoutant à l'équation (1) un terme ϵ_n représentant le bruit :

$$x_n = Wz_n + \mu + \epsilon_n \quad (3)$$

C'est cette dernière méthode qui est utilisé dans notre cas.

Toutes les méthodes d'analyse factorielle peuvent s'écrire comme une analyse en composantes principales (ACP) ou une décomposition en valeurs singulières d'un tableau de données particulier. L'ACP est donc au cœur de ces méthodes. L'approche classique pour gérer les données manquantes en ACP consiste à minimiser la fonction de coût (l'erreur de reconstitution) sur tous les éléments présents. Ceci peut être effectuer à travers un algorithme d'ACP itérative (aussi appelée expectation maximisation PCA, EM-PCA) décrit dans Kiers (1997). Celui-ci consiste à attribuer une valeur initiale aux données manquantes, effectuer l'analyse (ACP) sur le jeu rendu complet, compléter les données manquantes via la formule de reconstitution pour un nombre d'axes ou composantes fixé, et recommencer ces deux étapes jusqu'à convergence. Les paramètres (axes et composantes) ainsi que les données manquantes sont de cette manière simultanément estimées. Par conséquent cet algorithme peut être vu comme une méthode d'imputation simple. Il souffre cependant d'un problème de sur-ajustement qui peut être contourné grâce à une version régularisée de cet algorithme Josse and Husson (2010),Ilin and Raiko (2010).

Description des données utilisées

Les données utilisées sont en fait une combinaison entre les données de 07 stations professionnelles de la région ouest sur la période 1958-2000 (voir tableau 1) et les points des champs analysés de température sur 100 points de grille de ERA-40 sur la même période, couvrant un domaine s'étalant de -10^o à 12.5^o de longitude et de 17.5^o à 40^o de latitude. Ce domaine couvre toute l'Algérie et permet ainsi de couvrir un champ de température et de produire un nombre importants de points de grille (100) qui ont une influence sur la méthode d'imputation. En fait, chaque point de grille est considéré comme une station de mesure. En outre, comme les champs mensuelles sont disponibles que pour la température moyenne, il était impossible pour nous de les utiliser pour rattraper les températures minimales et maximales. Pour cela, nous avons produit des champs mensuels de la température minimale à partir de la moyenne des champs quotidiens du réseau de 06 heures, et les champs mensuels de la température maximale à partir de la moyenne des champs quotidiens de la température du réseau de 12 heures. On a admis ainsi, que la température minimale s'approchait de la température à 06 heures et que la température maximale s'approchait de la température à 12 heures.

Nous constituons une matrice de données sur la période 1958-2000 qui est le résultat d'une fusion entre les données des 07 stations professionnelles et les points de grilles. Les stations sont positionnées à l'intérieur de la matrice suivant leurs coordonnées géographiques.

Logiciels utilisés

L'ensemble du travail a été effectué sous R version 3.1.0 avec trois « packages » essentiels :

- La « package » reshape Wickham (2007) pour toute la gestion des données et leur mise en forme
- Le « package » ggplot2 Wickham (2009) pour la production des graphiques
- Et enfin, le « package » PCAMethods Stacklies et al. (2007) pour l'imputation des données manquantes

4. Stratégies pour la mise en oeuvre et pour la validation de la méthode

La disposition des données manquantes au sein de la matrice des données et leur pourcentage sont très importants dans le processus d'imputation des données manquantes. Plus leur pourcentage est faible et plus elles sont distribuées aléatoirement , plus le modèle convergera vers des estimateurs précis. Il faut rappeler une chose importante ici, est que ce processus ne donne pas la vraie valeur, mais seulement une estimation de cette valeur.

On distingue trois cas de distribution des valeurs manquantes au sein d'une matrice de données :

- La probabilité qu'une donnée soit manquante ne dépend ni des valeurs observées ni des valeurs manquantes (Missing completely at random (MCAR))
- La probabilité qu'une donnée soit manquante dépend des valeurs observées mais pas des valeurs manquantes (Missing at random (MAR))
- La probabilité qu'une donnée soit manquante dépend des valeurs manquantes (Missing not at random (MNAR))

Pour tester notre méthode sur nos données, nous choisissons une période ou la matrice est complète entre 1990-2000. Notre stratégie étant d'enlever aléatoirement ou non un certain pourcentage de données , dans la sous matrice constituée par les 07 stations professionnelles ,de cette matrice complète et de les ré-estimer par la méthode :

- En faisant varier le pourcentage des valeurs manquantes de 5% à 50% distribuées complètement aléatoirement (MCAR) (voir figure 1)
- Et enfin , en créant des blocs non aléatoires totalisant 24% de valeurs manquantes (MNAR)(voir figure 1).

On calcule à chaque fois une erreur relative moyenne donnée par l'expression suivante

$$ER = \frac{\Sigma_1^n (valeurobservee - valeurrattrappee)^2}{\Sigma_1^n (valeurobservee)^2} \quad (4)$$

Table 2. Algorithme d'imputation par ACP

Pour une matrice de données X incomplète, avec les éléments observés X_{obs} et les éléments manquants X_{mis} pour aboutir à une matrice complète X_{imp}, on réalise les étapes suivante.	
Etape 1	Initialiser les éléments manquants X_{mis} par les moyennes des lignes de X_{obs} $X_{mis} \leftarrow moyenne(X_{obs})$ pour donner matrice complète X_{imp}
Etape 2	Ré-estimer le terme biais en utilisant la matrice imputée X_{imp} par $\mu \leftarrow moyenne(X_{imp})$
Etape 3	Résoudre le système des k vecteurs propres de W par la méthode choisie, dans notre cas PPCA
Etape 4	Ré-estimation des valeurs manquantes par $X_{mis} \leftarrow WW^T(X_{imp}-\mu)+\mu$
Etape 5	Confirmer la convergence de X_{imp} ou $Z = W^T(X_{imp}-\mu)$. Si la convergence n'est pas atteinte, revenir à l'étape 2

n : étant le nombre de données manquantes

Ce sont les résultats de cette méthode de validation qui vont nous conforter sur le choix de la méthodologie adoptée, pour imputer les valeurs manquantes sur la période 1958-2000.

5. Mise en œuvre et résultats

Les résultats de la validation réalisée sur les températures mensuelles maximales puis minimales sur la période sans lacunes de 1990-2000 synthétisées dans le tableau 3 :

Table 3. Erreurs quadratiques moyennes en fonction des pourcentages des données manquantes pour les températures mensuelles maximales et minimales (1990-2000)

Fraction et Nbre de données manquantes	Erreur relative moyenne des T_{max}	Erreur relative moyenne des T_{min}
$5\% \rightarrow n = 46$	0.001912756	0.007389077
$10\% \rightarrow n = 92$	0.001464459	0.007404838
$15\% \rightarrow n = 138$	0.001342940	0.006791503
$20\% \rightarrow n = 184$	0.001352023	0.007041436
$25\% \rightarrow n = 231$	0.001403525	0.007102064
$30\% \rightarrow n = 277$	0.001577780	0.007274591
$50\% \rightarrow n = 462$	0.001870231	0.007914006

Il est évident que les résultats du test de validation effectué en créant artificiellement différents pourcentages de données manquantes d'une façon aléatoire sont excellents. En effet, l'erreur quadratique reste très faible de l'ordre de 0,007 même avec un pourcentage de valeurs manquantes de 50% !

Comme nos données manquantes ne sont pas disposées aléatoirement dans la période 1958-2000 (voir figure 2), il nous a apparu prudent d'estimer encore une fois l'erreur quadratique moyenne dans le cas ou les données sont retirées en blocs sur la matrice test de 1990-2000 (voir figure 1 en bas à droite),avec un pourcentage de 24%. Encore une fois, les résultats sont excellents puisque l'erreur quadratique moyenne s'est avérée très faible pour les températures maximales (égale à 0.0018) et pour les températures minimales (égale à 0.0061).

Confortés par ces résultats, nous avons appliquée l'imputation des données sur la période 1958-2000 pour combler les lacunes pour les 07 stations professionnelles. Mais cette fois-ci comme nous ne disposons pas des valeurs vraies pour calculer les erreurs quadratiques moyennes, nous nous sommes contentés à repérer le nombre de valeurs « anormales » comprises entre $\left[\overline{T}-2\sigma, \overline{T}+2\sigma\right]$ et $\left[\overline{T}-3\sigma, \overline{T}+3\sigma\right]$, à chaque fois $\overline{T}$ représente la température moyenne maximale ou minimale calculée sur la période 1990-2000 et σ son écart type.

Encore une fois, les résultats (voir tableau 4) sont très satisfaisants à excellents, car aucune valeur n'est vraiment anormale.

Table 4. Nombre de valeurs « anormales » et leur écart maximal par rapport à la normale des données imputées sur la période 1958-2000.

Intervalle	Température max. mensuelle	
	Nbre de valeurs	Ecart max. (^oC)
$> \overline{T}+2\sigma$	7	0,39
$< \overline{T}-2\sigma$	19	-1,67
$> \overline{T}+2\sigma$	0	/
$< \overline{T}-2\sigma$	0	/
Intervalle	**Température min. mensuelle**	
	Nbre de valeurs	Ecart max. (^oC)
$> \overline{T}+2\sigma$	4	0,77
$< \overline{T}-2\sigma$	3	-0,13
$> \overline{T}+2\sigma$	0	/
$< \overline{T}-2\sigma$	0	/

Toutes ces validations nous ont permis de reconstituer les séries des 07 stations sur la période de 1958 à 2013. Les résultats sont illustrés sur les graphiques 3 et 4.

6. Conclusions et perspectives

Les bons résultats obtenus dans cette modeste étude nous permettent d'ouvrir le grand chantier de la constitution des longues séries des principaux paramètres climatologiques dans notre pays. Ce chantier est devenu aujourd'hui possible grâce à la disposition du langage R et ses différents « packages » ainsi que des données des champs ré-analysés.

Dans le choix de la méthode d'imputation, notre choix s'est porté sur PCAmethods, méthode développée par des bio-informaticiens, malgré la disponibilité d'une dizaine d'autres méthodes sous R comme Amelia, MissMDA, mice ... qui sont techniquement plus précises (imputation multiple) , mais les erreurs quadratiques moyennes obtenues par la méthode utilisée nous ont surpris par leur faiblesse, qu'on peut considérer globalement comme négligeable.

Le travail n'est pas encore fini, car il reste maintenant la lourde tâche d'homogénéiser ces séries.

Il est très possible d'étendre cette étude à toute l'Algérie.

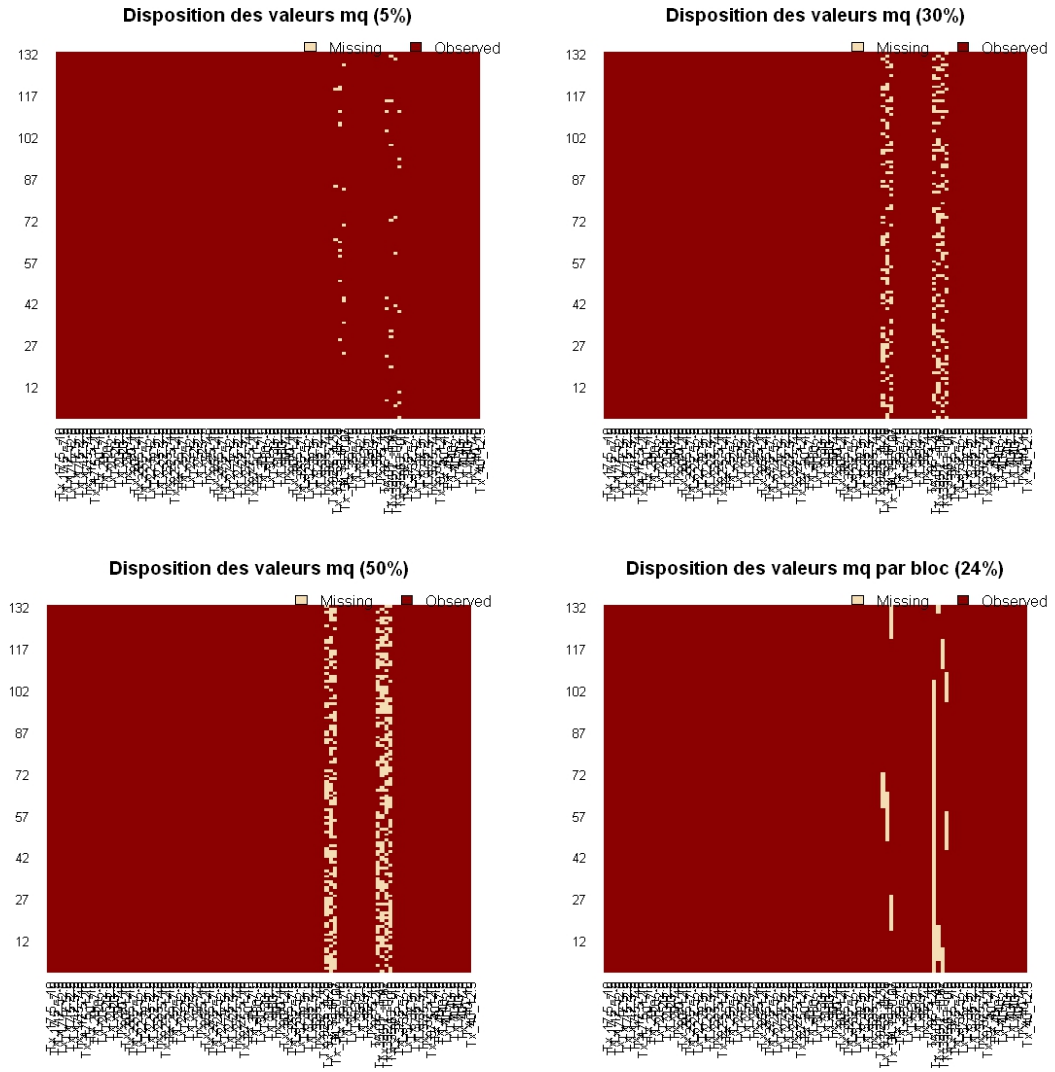

Figure 1. Disposition des valeurs manquantes au sein de la matrice de données test 1990-2000. L'axe X = stations+points de grilles et l'axe Y = chronologie (mois)

Remerciements

Ce travail est dédié à la mémoire de notre maitre et collègue monsieur Mohamed SENOUCI, qui nous a transmis sa passion de chercher.

Références

Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.*, 11 :1957–2000.

Josse, J. and Husson, F. (2010). Gestion des données manquantes en Analyse en Composantes Principales. In *Séminaire de laboratoire de mathématiques*, Bordeaux (FR), France.

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2) :251–266.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9) :1164.

Team, R. C. (2014). R : A language and environment for statistical computing.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612) :2961–3012.

Vatanen, T. (2012). Missing value imputation using subspace methods with applications on survey data. Master's thesis, Aalto university school of science.

Wickham, H. (2007). Reshaping data with the reshape packag. *Journal of Statistical Software*.

Wickham , H. (2009). *ggplot2 : elegant graphics for data analysis*. Springer.

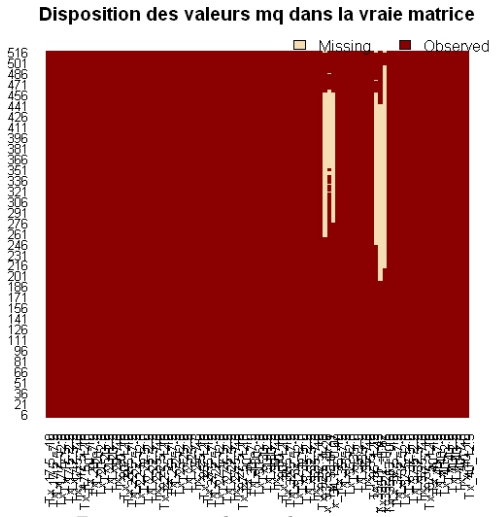

Figure 2. Disposition des valeurs manquantes au sein de la matrice de données 1958-2000. L'axe X = stations+points de grilles et l'axe Y = chronologie (mois)

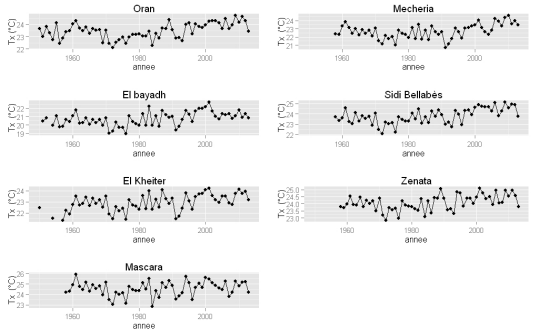

Figure 3. Evolution de la température maximale annuelle sur 07 stations de la région ouest

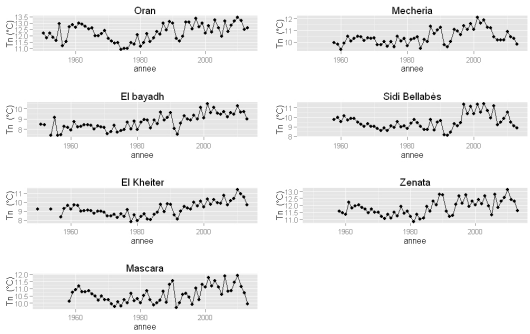

Figure 4. Evolution de la température minimale annuelle sur 07 stations de la région ouest