



LBRTI2101A

Analyse statistique de données spatiales et temporelles

**Projet d'analyse de la profondeur de la nappe phréatique
dans la région de Bakersfield, Californie**

Ayadi Youmna
Cleenewerck de Crayencour Harold
Dssam Abdelali
El Houda Teber Nour

Année académique 2024-2025

Table des matières

1	Introduction	2
2	Extraction des données	2
2.1	Importation des données et traitement des doublons	2
2.2	Filtrage temporel	3
3	Visualisation des valeurs de profondeur de la nappe	3
4	Analyse de la station 353539N1191118W001	4
5	Ajustement des données par rapport à l'altitude du sol	5
6	Analyse de la dépendance spatiale des résidus	8
7	Détection des valeurs extraordinaires via le krigeage	10
7.1	Comparaison du nombre de valeurs extraordinaires attendues et mesurées . .	10
7.2	Analyse des stations avec les valeurs aberrantes	11
8	Interpolation spatiale sans valeurs extraordinaires	13
9	Analyse des tendances de l'altitude de la nappe entre 2015 et 2021	15
A	Régression linéaire pour chaque année	18
A.1	Droites de régression linéaires	18
A.2	Distribution des résidus	20
A.3	Répartition spatiale de la profondeur de la nappe avant et après avoir retiré l'effet de l'altitude du sol	23
B	Variogrammes	25
C	Cartes de prédiction obtenues par krigeage	27
D	Programmes réalisés pour les différentes analyses	30

1 Introduction

Dans le cadre de ce projet, nous avons analysé les données de profondeur de la nappe phréatique dans la région de Bakersfield, en Californie, afin d'étudier les tendances spatiales et temporelles sur la période 2015-2021. Cette étude se concentre particulièrement sur les mois d'octobre, période clé marquant la fin de la saison des pluies.

L'objectif principal est de produire une carte des tendances de la profondeur de la nappe en utilisant des méthodes d'analyse statistique et géostatistique. Pour cela, nous avons sélectionné et traité les données pertinentes, modélisé les dépendances spatiales et effectué des prédictions afin de mettre en évidence les variations significatives et les anomalies.

Ce rapport détaille les étapes de l'analyse, les résultats obtenus et leurs interprétations.

2 Extraction des données

2.1 Importation des données et traitement des doublons

Dans la base de données initiale, nous avons identifié des observations partageant les mêmes coordonnées projetées (x, y), la même année et le même mois.

Ces doublons proviennent de stations géographiquement très proches, possédant des identifiants distincts (`site_code`), mais partageant les mêmes coordonnées projetées. Cela a conduit à l'enregistrement de plusieurs valeurs différentes pour la variable `mean_gse_gwe` sur une même position spatiale et pour une même période, ce qui constitue un problème pour les analyses géostatistiques, notamment le krigage.

Pour illustrer ce problème, nous présentons ci-dessous un tableau récapitulant certains des doublons détectés pour l'année 2015. Ces doublons ont été identifiés en regroupant les données selon leurs coordonnées projetées (x, y), l'année et le mois. Chaque doublon correspond à des stations différentes (`site_code`) partageant les mêmes coordonnées et la même période, tout en ayant des valeurs distinctes pour la variable `mean_gse_gwe`.

TABLE 1 – Extrait des données mesurées à la même période, avec les mêmes coordonnées (x, y), mais avec un numéro de code de station (`site_code`) et une mesure de la profondeur différentes (`mean_gse_gwe`).

year	month	x	y	longitude	latitude	site_code	mean_gse_gwe
2015	10	-2073181	1611193	-119.2138	35.26735	352673N1192138W001	163.8800
2015	10	-2073181	1611193	-119.2138	35.26735	352673N1192138W002	173.9300
2015	10	-2080397	1615552	-119.3032	35.28952	352895N1193032W001	174.0967
2015	10	-2080397	1615552	-119.3032	35.28952	352895N1193032W002	200.3967
2015	10	-2080397	1615552	-119.3032	35.28952	352895N1193032W003	263.7300

Pour résoudre ce problème, nous avons appliqué un processus de fusion des doublons. Chaque groupe de doublons, défini par une combinaison unique de ($x, y, \text{année}, \text{mois}$), a été

remplacé par une seule ligne.

La variable d'intérêt (`mean_gse_gwe`) a été calculée comme la moyenne des valeurs au sein de chaque groupe, tandis que le premier identifiant de station (`site_code`) ainsi que les coordonnées associées (longitude et latitude) ont été sélectionnés pour représenter le groupe.

2.2 Filtrage temporel

Nous avons sélectionné dans le jeu de données traité les observations correspondant aux mois d'octobre pour la période allant de 2015 à 2021. Cette sélection a été réalisée en filtrant les données sur les colonnes `month` et `year`.

3 Visualisation des valeurs de profondeur de la nappe

Après avoir retiré les doublons et sélectionné les données d'octobre 2015 à 2021 dans le jeu de donnée original, nous présentons en figure 1 une carte des valeurs moyennes de profondeur de la nappe en octobre de 2015 à 2021 aux différentes stations de mesure. Cette carte indique également l'élévation du sol dans la région. Ces données d'élévation du sol proviennent du fichier `DEM_grid.csv`, qui est un modèle numérique de terrain.

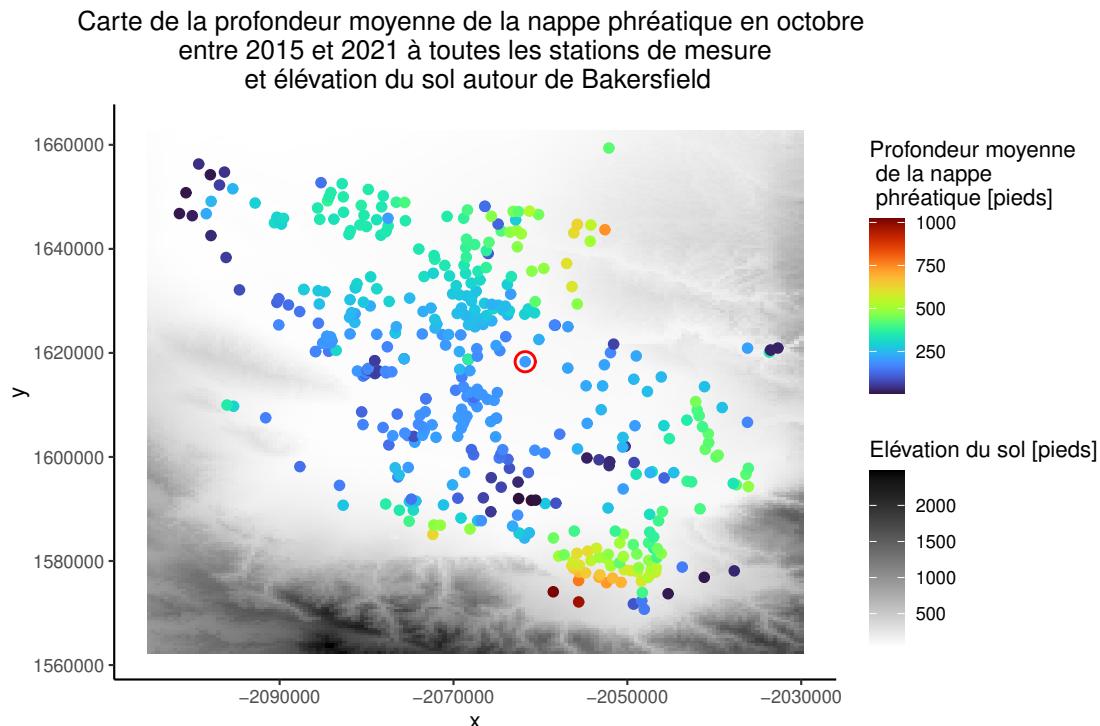


FIGURE 1 – Carte de la profondeur moyenne de la nappe phréatique en octobre entre 2015 et 2021 à toutes les stations de mesure (points colorés) et élévation du sol (en gris) dans la région de Bakersfield. La station **353539N1191118W001** est entourée en rouge.

En observant cette carte, une corrélation positive entre la profondeur de la nappe et l'élévation du sol semble exister. En effet, lorsque l'élévation du sol est grande (en bas de la carte), la profondeur de la nappe phréatique est en moyenne plus élevée. A l'inverse, les zones sur la carte où l'altitude est plus faible semblent être associées à une profondeur moyenne de la nappe plus faible (centre et nord-ouest sur la carte). Il n'est donc pas raisonnable de considérer que l'espérance de la profondeur de la nappe est constante sur la zone d'intérêt et il n'y a donc pas de stationnarité d'ordre 1. L'effet de l'altitude sur la profondeur de la nappe sera retiré par régression linéaire dans la section 5.

La station **353539N1191118W001** a été mise en évidence sur cette carte. L'évolution de la profondeur la nappe phréatique au niveau de cette station, entre décembre 1946 et février 2024, est analysée dans la section suivante de ce rapport.

4 Analyse de la station 353539N1191118W001

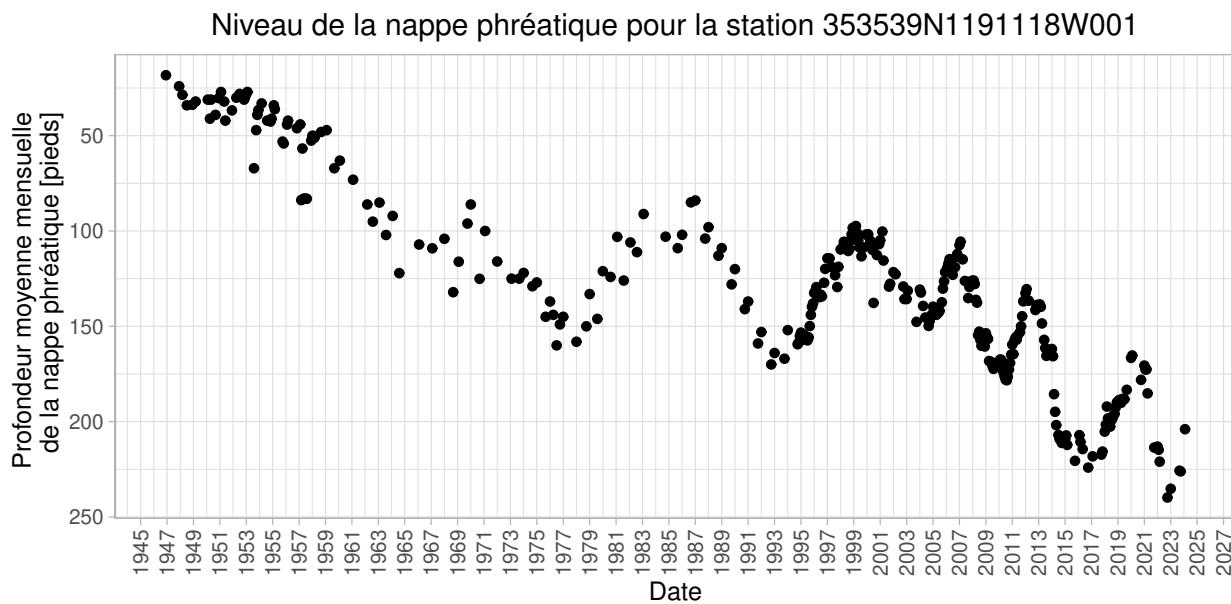


FIGURE 2 – Série temporelle du niveau de la nappe phréatique pour la station **353539N1191118W001**

On observe une tendance générale à la baisse du niveau de la nappe phréatique pour la station étudiée. Cette évolution se décompose en trois phases distinctes :

De 1945-1979 : une diminution marquée et continue du niveau de la nappe phréatique passant de 25 à 150 pieds. Cette baisse pourrait être due au développement massif de l'irrigation agricole d'après la deuxième guerre mondiale ([Bierkens and Wada, 2019](#)).

De 1980-2000 : une période caractérisée par une relative stabilisation, oscillant entre 100 et 150 pieds. On observe des cycles de recharge et de décharge, suggérant un équilibre précaire. Cela peut être dû à la modernisation des systèmes d'irrigation et la diversification des sources d'approvisionnement en eau.

Après 2000 : la tendance à la baisse reprend, particulièrement accentuée après 2015, où le niveau atteint des profondeurs historiques entre 200 et 250 pieds. Bien que des oscillations saisonnières sont observées, le niveau de la nappe ne parvient jamais à retrouver son état initial, indiquant que les prélèvements sont supérieurs à la capacité de recharge naturelle. Cette aggravation peut être attribuée à plusieurs facteurs convergents : l'insuffisance des précipitations due au changement climatique, la pression démographique croissante, et la multiplication des épisodes de sécheresse. Entre 2012 et 2016, la Californie a subit une sécheresse qui a causé l'épuisement des eaux souterraines, l'assèchement de puits, et une pénurie d'eau dans plus de 2000 foyers ([Escriva-Bou et al., 2020](#)).

5 Ajustement des données par rapport à l'altitude du sol

Comme mentionné dans la section 3, la profondeur de la nappe phréatique semble dépendre de l'altitude du sol. Comme cette altitude varie considérablement sur la zone d'étude, il ne semble pas y a voir stationnarité d'ordre 1. Pour garantir cette stationnarité d'ordre 1, nous avons donc retiré l'effet de l'élévation du sol à l'aide d'un modèle de régression linéaire simple. La fonction espérance de la profondeur de la nappe phréatique s'écrit ainsi :

$$\mu(\mathbf{x}) = \beta_0 + \beta_1 g_1(\mathbf{x})$$

Où $g_1(\mathbf{x})$ représente l'élévation du sol, qui dépend des coordonnées \mathbf{x} . Grâce à ce modèle de régression linéaire, nous pourrons travailler avec les résidus $\epsilon(\mathbf{x}) = Z(\mathbf{x}) - \mu(\mathbf{x})$ dont l'espérance est nulle pour tout \mathbf{x} . $Z(\mathbf{x})$ est la profondeur de la nappe phréatique.

Avant de pouvoir estimer les paramètres β_0 et β_1 de cette équation, il est important de remarquer que les stations de mesures ne se trouvent pas exactement aux mêmes points (x, y) que les points de mesure de l'élévation du sol (`DEM_grid`). Il est donc nécessaire de prédire l'altitude du sol aux stations de mesure. Pour ce faire, la méthode de prédiction linéaire de tessellation de Thiessen / Voronoï a été utilisée. Cette méthode a été choisie grâce à la simplicité de son implémentation en R. Bien que la performance de cette méthode soit souvent moins bonne que celle des autres méthodes de prédiction linéaire couramment utilisées, ceci ne devrait pas poser de problème car la grille `DEM_grid` contient énormément de points très rapprochés les uns des autres sur la zone d'étude, la résolution étant de 500m. Par conséquent, considérer que l'altitude du sol au niveau d'une station de mesure est égale à l'altitude du point de `DEM_grid` le plus proche n'est probablement pas une mauvaise approximation.

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont estimés par la méthode des moindres carrés ordinaire (méthode OLS).

Pour octobre 2015, l'équation de la droite de régression estimée par méthode OLS est égale à :

$$\mu(\mathbf{x}) = -31.2950 + 2.5943g_1(\mathbf{x})$$

Et cette droite de régression linéaire est tracée sur la figure 3.

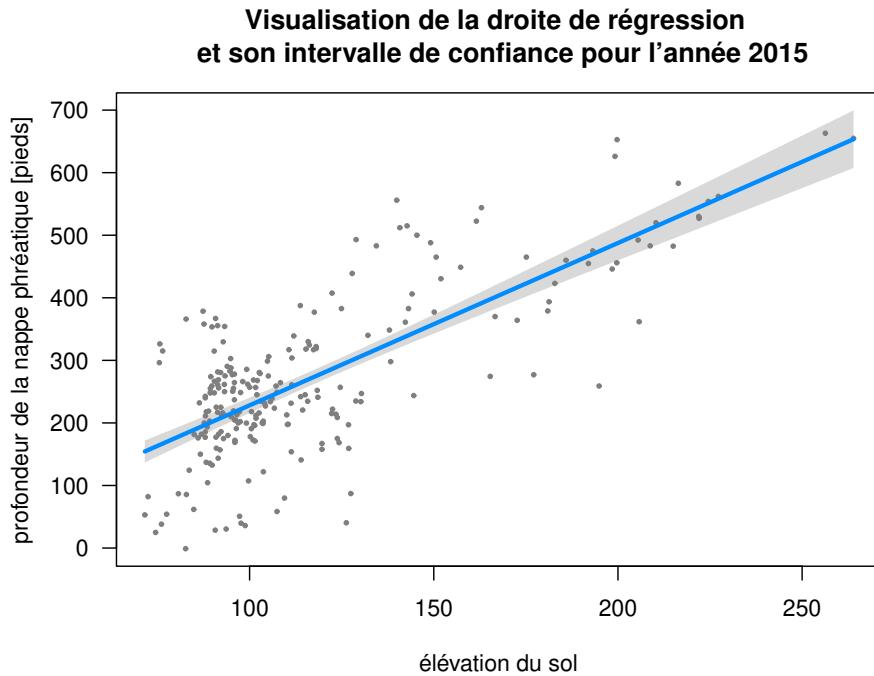


FIGURE 3 – Droite de régression linéaire et son intervalle de confiance pour octobre 2015. La moyenne mensuelle de la profondeur de la nappe phréatique est en pieds

Le R^2 , qui quantifie la proportion de la variance totale expliquée par le modèle et donc la qualité de l'ajustement, vaut 55.75%. De plus, les résidus semblent assez bien suivre une distribution normale, comme le montre l'histogramme des résidus et le Q-Q plot (cf. figure 4). Les droites de régression linéaires, les valeurs des estimateurs de β_0 et β_1 , et les valeurs des R^2 pour chaque année se trouvent dans l'annexe A.

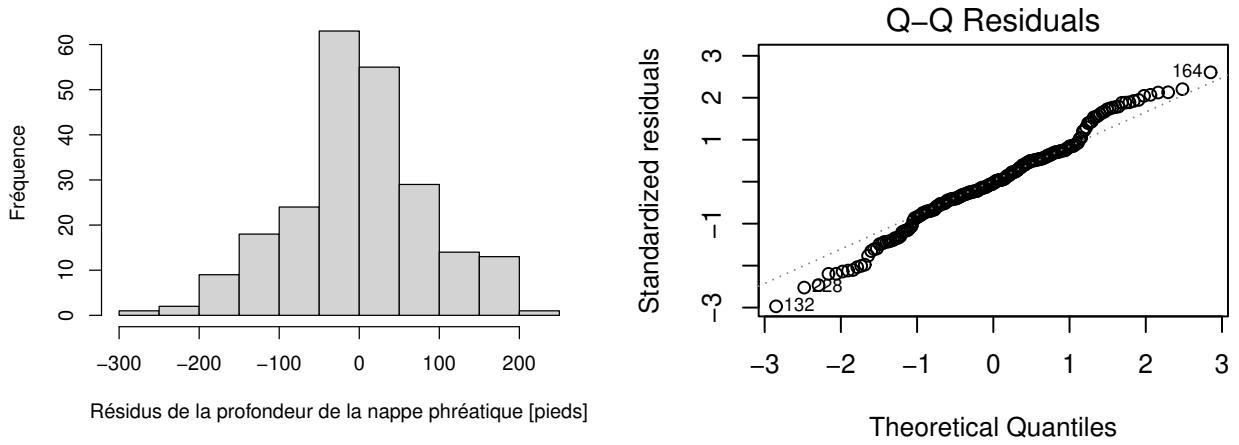


FIGURE 4 – A gauche : histogramme des résidus de la profondeur de la nappe phréatique (en pieds) en octobre 2015 obtenus après avoir retiré l’effet de l’élévation du sol par régression linéaire. A droite : Q-Q plot de comparaison de la distribution observée des résidus standardisés de la profondeur de la nappe avec une loi normale

Bien que la qualité de l’ajustement de la droite de régression linéaire soit assez bonne pour les données d’octobre 2015, ce n’est pas le cas pour toutes les années (cf. annexe A). En effet, pour octobre 2021, le modèle n’explique que 20.2% de la variance totale ($R^2 = 20.2\%$, cf. tableau 6). En outre, les résidus obtenus après avoir retiré l’effet de l’élévation ne suivent pas une distribution normale pour toutes les années, notamment pour 2017 et 2019 (cf. figures 12d et 12h). Or, la détection des valeurs extraordinaires par la méthode du Leave-One-Out Cross-Validation plus tard dans ce rapport suppose que les résidus suivent une loi normale.

Cependant, pour garder l’analyse des données pour chaque année uniforme, nous estimons la fonction espérance par régression linéaire pour chaque année et nous considérons que les résidus suivent une distribution normale pour chaque année, même si cette approximation n’est pas très bonne pour octobre 2017 et 2019.

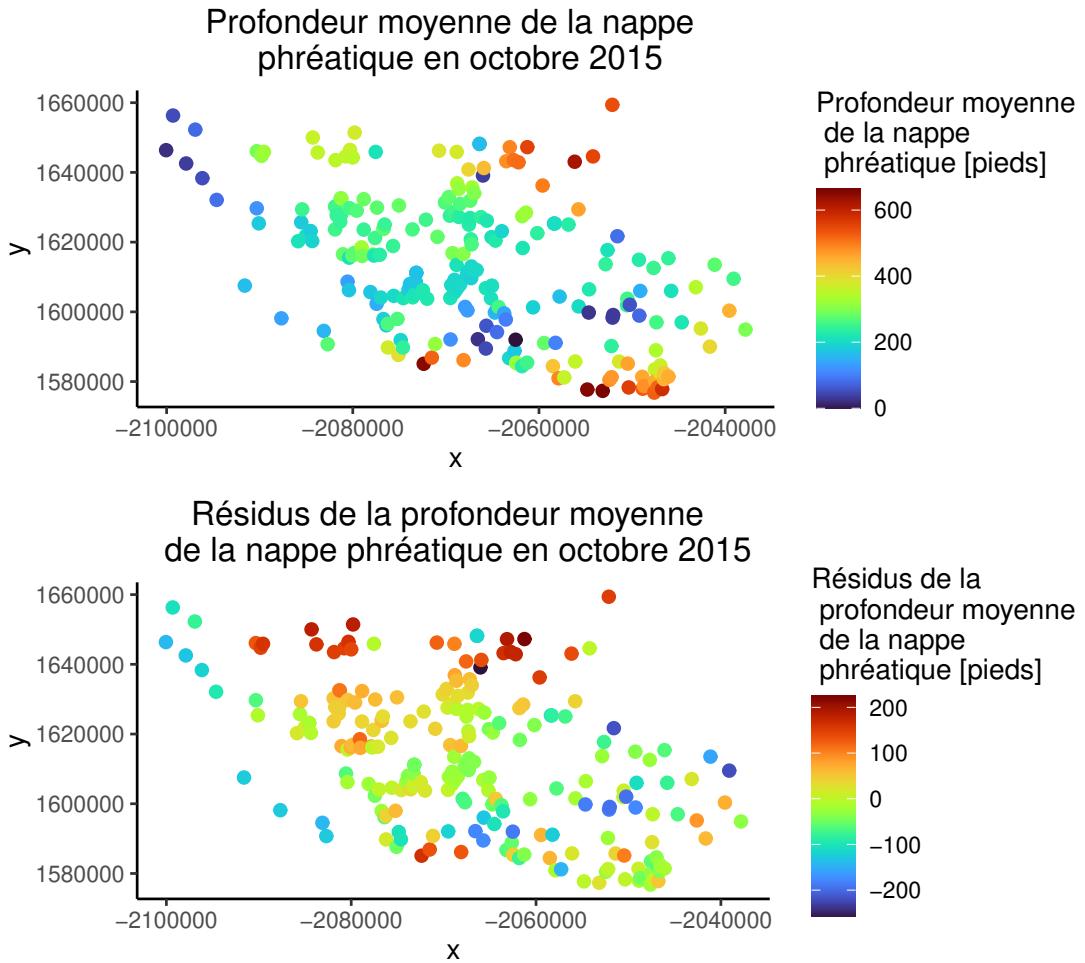


FIGURE 5 – Carte de la profondeur (en haut) et des résidus de la profondeur (en bas) moyenne de la nappe phréatique en octobre 2015 à toute les stations de mesure autour de Bakersfield

Les résidus de la profondeur moyenne de la nappe phréatique en octobre 2015, visibles sur la carte en figure 5, varient entre -200 et 200 pieds et ces résidus sont proches de 0 au niveau de nombreuses stations. Les valeurs élevées de profondeur aux endroits où l’élévation du sol est élevée (au sud-est de la carte) ne sont plus visibles, tout comme les valeurs basses au centre et au nord-ouest de la carte, là où l’altitude est plus faible. Avoir retiré l’effet de l’élévation du sol sur la profondeur de la nappe a donc permis de se rapprocher de la stationnarité d’ordre 1.

6 Analyse de la dépendance spatiale des résidus

Nous avons ensuite analysé la dépendance spatiale des résidus. Pour ce faire, il est nécessaire d’utiliser des variogrammes expérimentaux. Ceux-ci permettent de choisir les modèles de variogrammes appropriés (modèle exponentiel, sphérique, gaussien ou effet pépite) et de potentiellement choisir de combiner plusieurs modèles. Les variogrammes expérimentaux

permettent également d'estimer les paramètres du modèle de variogramme.

La distance maximale utilisée pour créer ces variogrammes expérimentaux a été fixée à un tiers de la distance maximale entre les stations de mesure. Cependant, pour certaines années, cette distance a été réduite pour améliorer l'ajustement des modèles aux variogrammes expérimentaux.

Si l'on souhaite obtenir des variogrammes corrects, il est nécessaire d'avoir une stationnarité d'ordre 1, c'est-à-dire que la valeur de l'espérance reste constante à tout endroit de la zone d'étude. Cette condition a été respectée grâce à la régression linéaire réalisée plus tôt.

Afin de créer les variogrammes pour le mois d'octobre de chaque année, il a été décidé de combiner un modèle exponentiel et un effet pépite (“nugget”). Les valeurs optimisées des paramètres des modèles ont été obtenues par la méthode des moindres carrés. Pour le variogramme modélisé de 2015, ces valeurs sont reprises dans le tableau ci-dessous. Les variogrammes obtenus pour les autres années et les valeurs de paramètres correspondants se trouvent en annexe B.

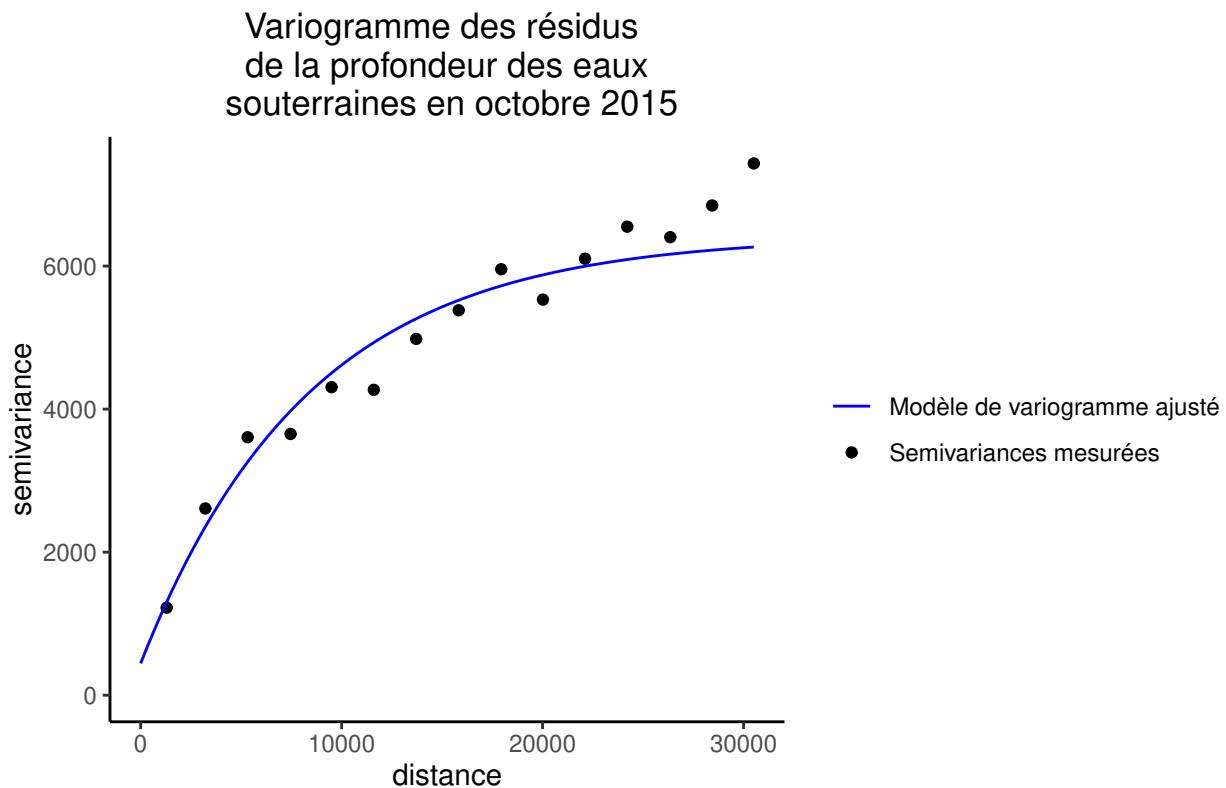


FIGURE 6 – Variogramme expérimental et modélisé des résidus de la profondeur des eaux souterraines en octobre 2015

TABLE 2 – Valeurs optimisées des paramètres du modèle de variogramme obtenu à partir des données de profondeur de la nappe en octobre 2015

	Palier $\hat{\sigma}^2$	Portée ($\hat{\theta}$)
Pépite	443.7867	0
Exponentiel	5976.7766	8350.166

Il est important de noter que les variogrammes modélisés ne s'ajustent pas toujours bien aux variogrammes expérimentaux correspondants. C'est notamment le cas pour 2016 et, en particulier, 2018.

7 Détection des valeurs extraordinaires via le krigeage

7.1 Comparaison du nombre de valeurs extraordinaires attendues et mesurées

Après avoir modélisé les variogrammes pour chaque année, nous allons les utiliser pour effectuer un krigeage avec une validation croisée de type LOOCV (Leave-One-Out Cross Validation) afin de détecter les valeurs extraordinaires qui sortent de l'intervalle de prédiction avec un niveau de confiance de 99 %. Autrement dit, nous allons réaliser le krigeage autant de fois qu'il y a de lignes dans le jeu de données pour le mois d'octobre d'une année entre 2015 et 2021, en excluant une ligne à chaque fois pour la prédire par krigeage. Ensuite, nous vérifions, à l'aide de la valeur et de la variance prédictes, si la valeur observée tombe dans l'intervalle de confiance et calculons sa p-valeur. Cette méthode est appliquée pour chaque année afin d'obtenir une liste de ces valeurs avec leurs p-valeurs.

Après cela, nous avons refait le krigeage sur toute la grille de prédiction, en excluant les valeurs marquées comme extraordinaires. Les valeurs extraordinaires observées en octobre 2015 et leurs p-valeurs associées sont présentées sur la figure 8. Les cartes obtenues pour les autres années se trouvent dans l'annexe C.

Le tableau 3 reprend le nombre de valeurs extrêmes attendues et observées pour chaque année.

TABLE 3 – Comparaison du nombre total de valeurs extraordinaire identifiées pour l'ensemble des mois d'octobre entre 2015 et 2021 avec le nombre attendu par la méthode LOOCV

Année	Valeurs extrêmes attendues	Valeurs extrêmes observées
2015	2.29	10
2016	1.6	7
2017	2.84	19
2018	2.47	22
2019	2.19	20
2020	2.31	4
2021	2.23	11

Nous constatons que pour l'année 2015, 10 valeurs considérées comme extraordinaire sont obtenues. Or, par définition, nous nous attendons en moyenne à ce qu'environ 1 % des stations soient marquées comme extraordinaire. Dans notre cas, cela correspond à environ $229 \cdot 0,01 \approx 3$ points. En réalité, nous soupçonnons que cette différence soit due au fait que les hypothèses utilisées pour la détection des valeurs aberrantes ne sont pas toutes parfaitement remplies. En effet, bien que les résidus soient normaux pour certaines années (confirmé par le test de Shapiro-Wilk avec une p-valeur de 0,033 pour 2015), ce n'est pas le cas pour toutes les années (voir section 5). De plus l'hypothèse de stationnarité d'ordre 1 pourrait ne pas être totalement satisfaite comme semblent l'indiquer les faibles R^2 de certaines droites de régression et le mauvais ajustement des variogrammes modélisés par rapport aux variogrammes expérimentaux. Par conséquent, le krigeage sous-estime la variance et prédit des intervalles de confiance trop petits, ce qui entraîne une détection excessive de points considérés comme extraordinaire.

De plus, pour l'année 2020, nous disposons du meilleur variogramme expérimental, c'est-à-dire celui pour lequel il semble y avoir le moins de fluctuations locales / de bruit (cf. figure 14e), et c'est également l'année où le nombre de valeurs détectées est le plus proche du nombre des valeurs attendues.

7.2 Analyse des stations avec les valeurs aberrantes

En observant les stations considérées comme extraordinaire, on remarque qu'il existe des stations détectées comme aberrantes pendant plus d'une année. Par exemple, une seule station, **353072N1188037W001**, a été identifiée comme aberrante durant 5 années consécutives, de 2017 à 2021, avec des profondeurs moyennes de 175,8 pieds et un écart type de 13,2 pieds. Ces valeurs, bien que stables, sont tout de même considérées comme extraordinaire.

Nous avons également constaté que la station **352015N1192094W001** est la seule présentant une profondeur de la nappe phréatique de -285,7 pieds, ce qui indique probablement une erreur de mesure. En effet, dans l'ensemble des données originales, seules trois stations affichent des valeurs de profondeurs très négatives. Cela pourrait être dû à un dysfonctionnement de l'appareil de mesure pour ces trois cas. Nous pensons qu'avoir de petites valeurs

négatives de profondeur peut survenir, probablement en raison d'une inondation ou d'une situation particulière. Cependant, des valeurs négatives aussi importantes semblent suspectes et probablement erronées.

TABLE 4 – Six premières lignes du jeu de données monthly_mean delimited.csv, ordonnées par ordre décroissant des profondeurs (mean_gse_gwe)

site_code	year	month	mean_gse_gwe	x	y	longitude	latitude
355725N1193941W001	2013	3	-304.9	-2080636	1648265	-1193941	3557248
352015N1192094W001	2016	10	-285.7	-2074577	1603944	-1192094	3520148
355506N1195271W001	2012	2	-227.2	-2092822	1648814	-1195271	3555064
355944N1195814W001	2014	2	-42.5	-2096347	1654761	-1195814	3559441
355911N1196357W001	2006	1	-3.0	-2101164	1655600	-1196357	3559110
353277N1193257W001	2012	10	-2.9	-2081330	1620191	-1193257	3532770

Ensuite, nous avons réalisé une représentation graphique des stations marquées comme étant aberrantes au cours des différentes années, en indiquant le nombre d'années durant lesquelles elles ont été flaguées comme extraordinaires, comme illustré dans la figure ci-dessous (figure 7) :

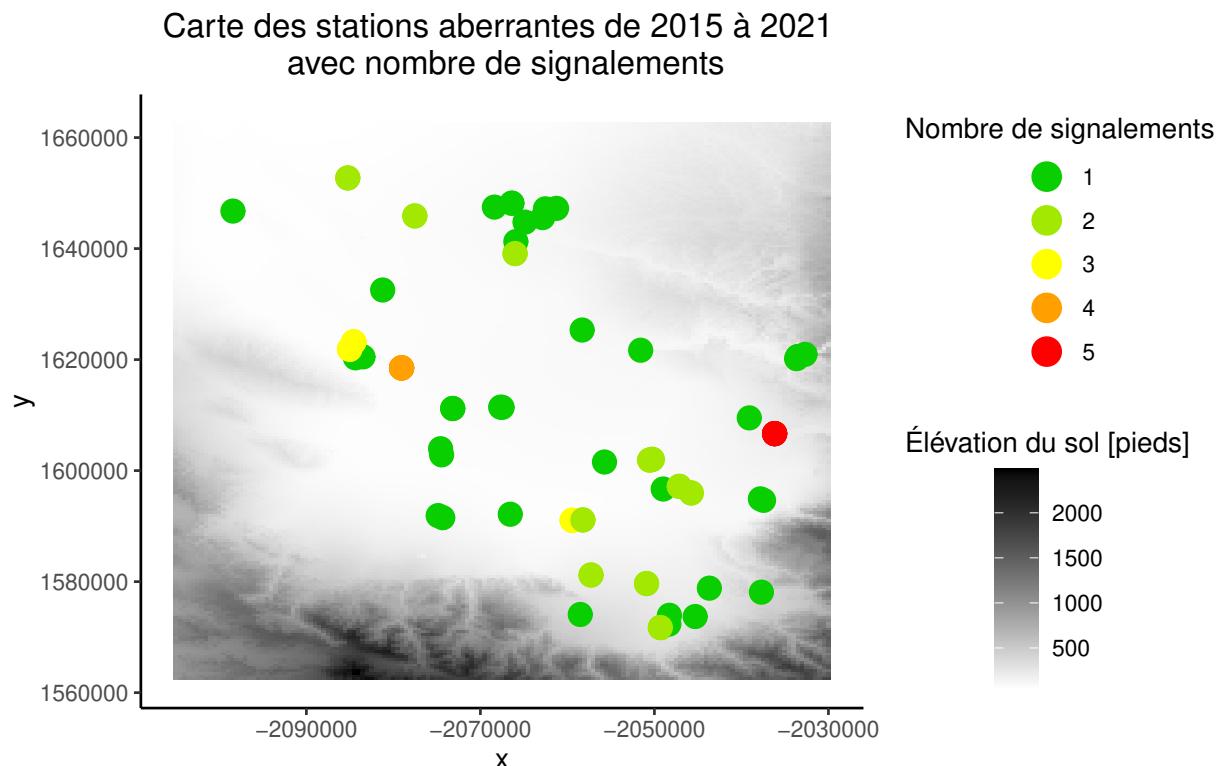


FIGURE 7 – Stations marquées comme aberrantes entre 2015 et 2021 et nombre de signalements

Nous constatons que la majorité des outliers n'ont été signalés qu'une seule fois. Pour

les stations ayant été signalées plusieurs fois au cours des sept années consécutives, celles-ci ne semblent pas se concentrer dans une zone géographique particulière. Cette répartition indique qu'aucune tendance claire ne se dégage quant à l'apparition des outliers, ni dans leur position géographique, ni dans leur fréquence.

8 Interpolation spatiale sans valeurs extraordinaires

Après avoir appliqué la validation croisée LOOCV (Leave-One-Out Cross Validation) sur le jeu de données original pour détecter les valeurs extraordinaires sortant de l'intervalle de confiance, les valeurs identifiées comme aberrantes ont été retirées pour garantir une meilleure qualité des prédictions.

Le jeu de données nettoyé a été utilisé pour réaliser le krigage de la profondeur de la nappe phréatique pour chaque mois d'octobre entre 2015 et 2021. La carte de krigage correspondant à l'année 2015 est présentée ci-dessous. Les autres cartes se trouvent dans l'annexe (cf. figures 15a à 15f).

Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2015 et points aberrants avec leur p-valeurs associées

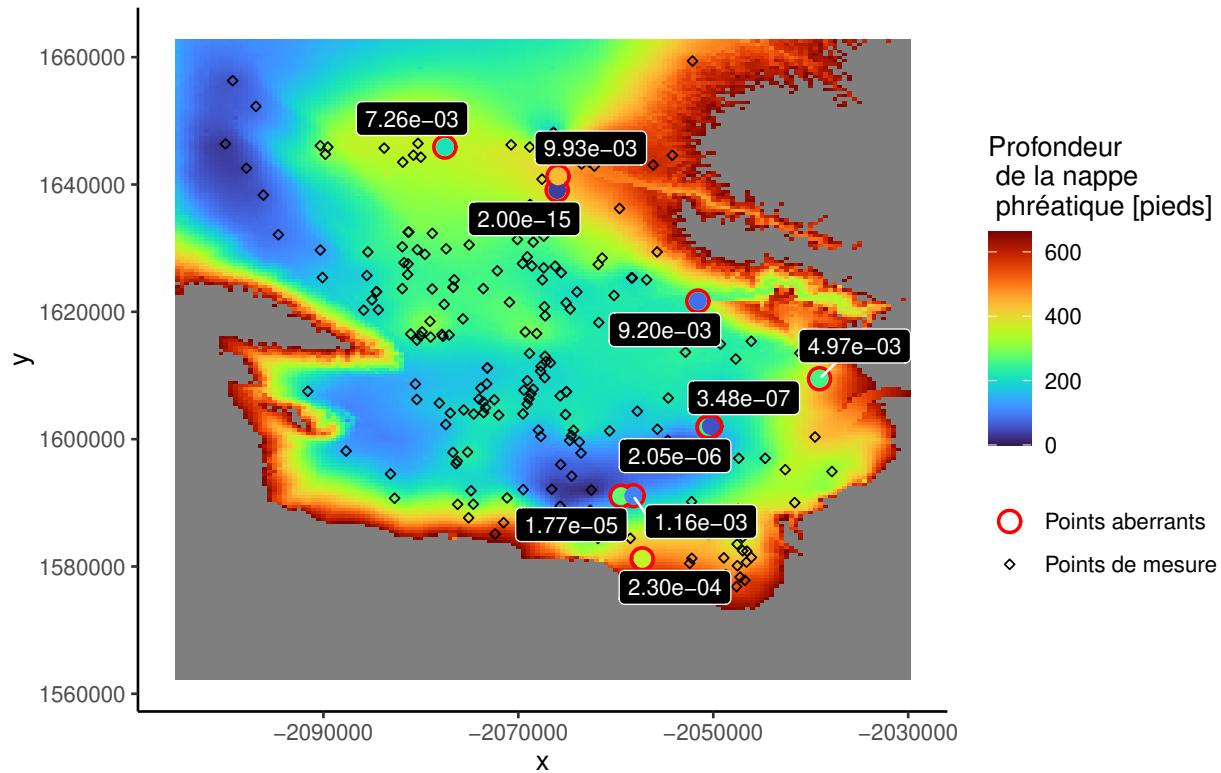


FIGURE 8 – Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2015 et points aberrants avec leur p-valeurs associées. Les valeurs de profondeur prédites dépassant la profondeur maximale observée ne sont pas représentées sur la carte.

D'après cette figure, on trouve que les points aberrants se trouvent principalement dans des zones de transition spatiale marquée, où les prédictions de la nappe changent rapidement entre des valeurs élevées et faibles.

En comparant ces valeurs aberrantes par rapport aux points de mesures à proximité, on remarque que les zones centrales présentent une cohérence élevée entre les prédictions et les observations ce qui se traduit par des valeurs aberrantes moins fréquentes, alors que les zones périphériques et celles à fort gradient spatial montrent plus de points aberrants, ce qui est peut-être dû au manque de stations de mesure dans ces régions.

Pour chaque pixel de la grille d'interpolation, nous avons calculé la moyenne des variances de prédiction issues du krigage pour les mois d'octobre de 2015 à 2021 afin d'identifier les zones où l'incertitude des prédictions est faible ou élevée sur l'ensemble de la période étudiée (cf. figure 9).

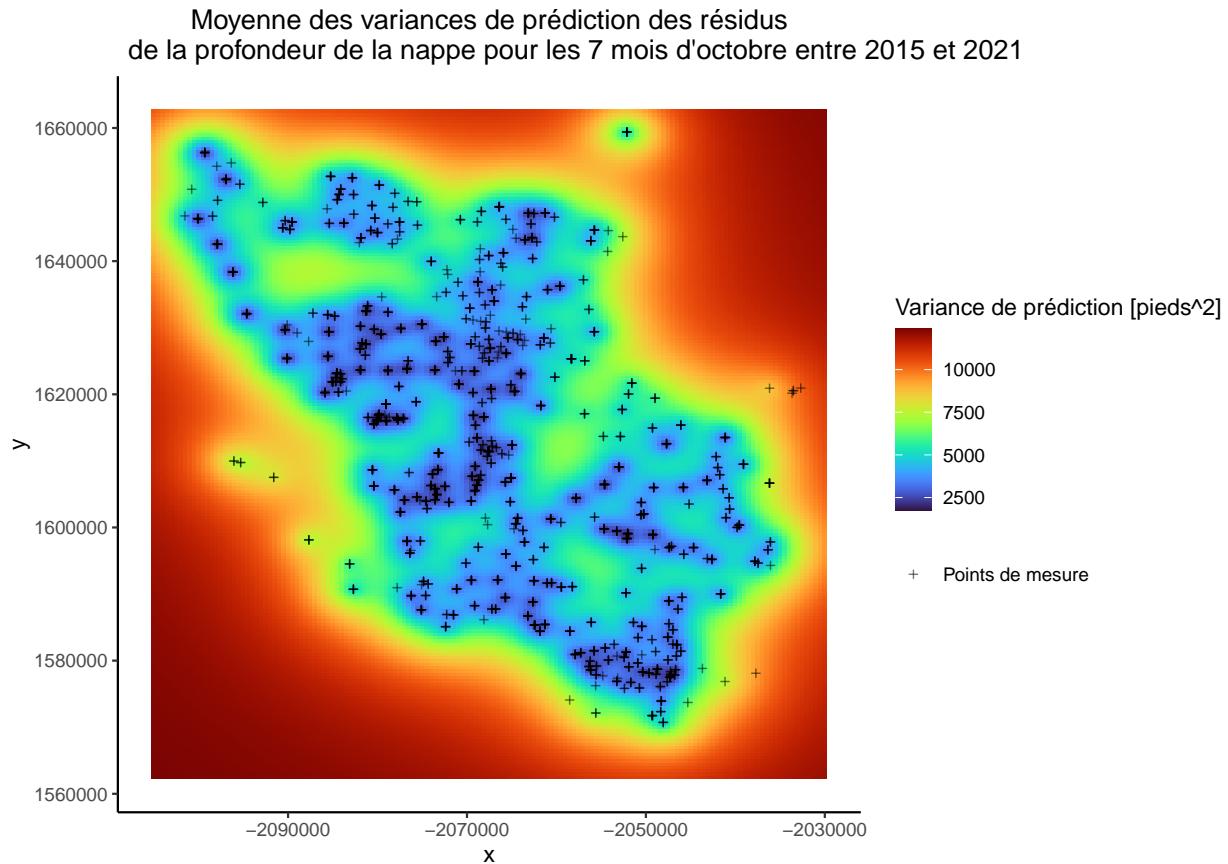


FIGURE 9 – Moyenne des variances de prédiction des résidus de la profondeur de la nappe pour les 7 mois d'octobre entre 2015 et 2021. Tous les points de mesure sont marqués sur cette carte bien que la profondeur de la nappe en octobre n'ait pas été mesurée chaque année pour certaines des stations.

D'après cette figure, on trouve que les zones à variance de prédiction élevée sont situées en périphérie et dans des régions moins densément couvertes par des points de mesure. Cela est principalement dû au manque de données dans ces régions, ce qui a entraîné des prédictions plus incertaines par le krigage. Plus la distance par rapport aux points de mesure augmente,

plus la variance de prédiction se rapproche de la variance des résidus. Au niveau du centre de la carte, les variances sont faibles grâce à une forte densité des points de mesure, ce qui permet des prédictions plus fiables. Les valeurs aberrantes y sont rares et souvent isolées.

9 Analyse des tendances de l'altitude de la nappe entre 2015 et 2021

Après avoir estimé la profondeur de la nappe phréatique pour chaque mois d'octobre entre 2015 et 2021, nous pouvons nous demander comment cette profondeur a évolué au cours des années dans la région de Bakersfield. Le niveau de la nappe aurait-il baissé sur l'ensemble de la région en raison du réchauffement climatique, à cause des sécheresses de plus en plus fréquentes, ou à cause de l'utilisation trop intense de cette eau pour l'approvisionnement de la population croissante, l'agriculture et les industries ([Kang and Jackson, 2016](#)) ? Ou au contraire, le niveau de la nappe aurait-il augmenté grâce à des mesures de protection et de gestion plus durable, notamment grâce au "*Sustainable Groundwater Management Act (SGMA)*" adopté en 2014 ([Escriva-Bou et al., 2020](#)) ? Il est également possible que le niveau de la nappe n'ait pas connu de variations significatives au cours de ces sept années.

Pour répondre à ces questions, nous avons réalisé une carte des tendances : pour chaque pixel de la carte de prédiction, la pente d'un modèle de régression linéaire simple a été déterminée. Ce modèle permet de prédire l'évolution de la profondeur de la nappe en fonction de l'année. Chaque pente a été déterminée par la méthode des moindres carrés ordinaires.

Ensuite, il est nécessaire d'évaluer la significativité de chacune de ces pentes. Pour cela, nous avons réalisé un test de Student afin de vérifier si chaque pente est statistiquement significativement différente de zéro. L'hypothèse nulle, selon laquelle la pente obtenue par régression linéaire est nulle, est rejetée si la p-valeur associée au test est inférieure à 5%.

La carte des tendances est présentée sur la figure 10. La tendance n'est pas significativement différente de 0 sur la majorité de la grille de prédiction. Toutefois, plusieurs régions révèlent une baisse du niveau de la nappe (augmentation de la profondeur de la nappe entre 0 et 20 pieds) tandis que d'autres régions présentent une légère augmentation du niveau de la nappe.

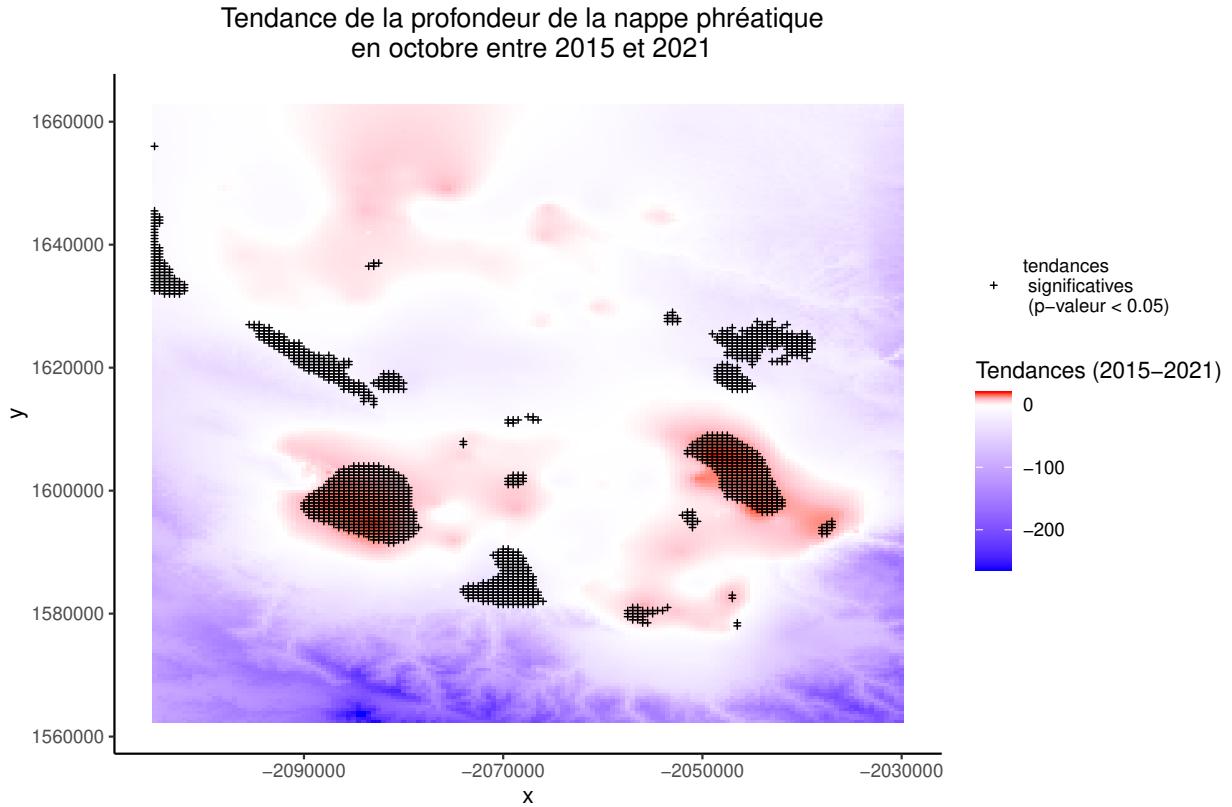


FIGURE 10 – Carte des tendances de la profondeur de la nappe phréatique en octobre entre 2015 et 2021

Le niveau de la nappe semble avoir principalement baissé dans deux zones sur la carte (autour de $y = 1600000$). Sur les cartes obtenues par krigeage entre 2015 et 2021, ces deux zones correspondent à des régions où la nappe est peu profonde (indiquées par les zones bleues). Nous pouvons donc émettre l'hypothèse que l'eau souterraine dans ces deux zones est davantage utilisée, notamment pour la consommation ou l'agriculture, car elle est plus facile d'accès par rapport aux nappes plus profondes.

A l'inverse, les zones où le niveau de la nappe a légèrement augmenté se situent principalement à proximité de zones où l'élévation du sol est élevée et, par conséquent les nappes sont plus profondes. Ces nappes plus profondes sont probablement moins attrayantes pour l'installation de puits, et elles sont donc probablement moins perturbées par l'activité humaine. Ainsi, ces zones ont peut-être pu accumuler plus d'eau au fil des années.

Enfin, au sud de la carte, la tendance semble être très négative, ce qui correspondrait à un niveau croissant de la nappe dans cette zone. Cependant, la tendance pour aucun de ces pixels n'est statistiquement significative. Nous ne pouvons donc pas conclure que la tendance est différente de zéro. Il est également important de souligner que les prédictions de profondeur pour toutes les années, et donc les prédictions des tendances dans cette zone, sont très mauvaises. En effet, il y a une accumulation d'erreurs provenant de la régression linéaire (profondeur de la nappe en fonction de l'élévation du sol), des modèles de variogrammes, et des variances de prédiction du krigeage.

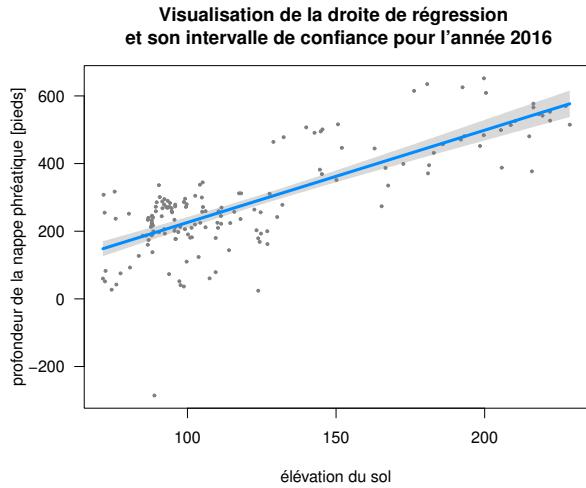
En ce qui concerne la régression linéaire, lorsque l'élévation du sol est élevée, la droite de régression ne s'ajuste pas bien aux points de mesure, et la profondeur des nappes varie énormément à ces altitudes. Par conséquent, lorsque l'espérance de la profondeur est ajoutée aux résidus après le krigage, la prédiction est particulièrement mauvaise pour les zones de la carte où l'élévation est élevée. De plus, l'ajustement parfois mauvais des modèles aux variogrammes expérimentaux a été discutée précédemment. Finalement, dans les zones d'extrapolation, loin des stations de mesure, la variance de prédiction augmente rapidement, ce qui montre que la prédiction est de mauvaise qualité dans ces régions.

Références

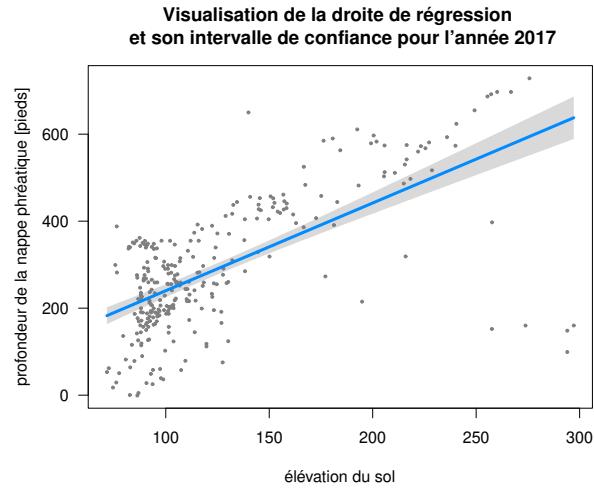
- Bierkens, M. F. and Wada, Y. (2019). Non-renewable groundwater use and groundwater depletion : a review. *Environmental Research Letters*, 14(6) :063002.
- Escriva-Bou, A., Hui, R., Maples, S., Medellín-Azuara, J., Harter, T., and Lund, J. (2020). Planning for groundwater sustainability accounting for uncertainty and costs : An application to California's central valley. *Journal of environmental management*, 264 :110426.
- Kang, M. and Jackson, R. B. (2016). Salinity of deep groundwater in California : Water quantity, quality, and protection. *Proceedings of the National Academy of Sciences*, 113(28) :7768–7773.

A Régression linéaire pour chaque année

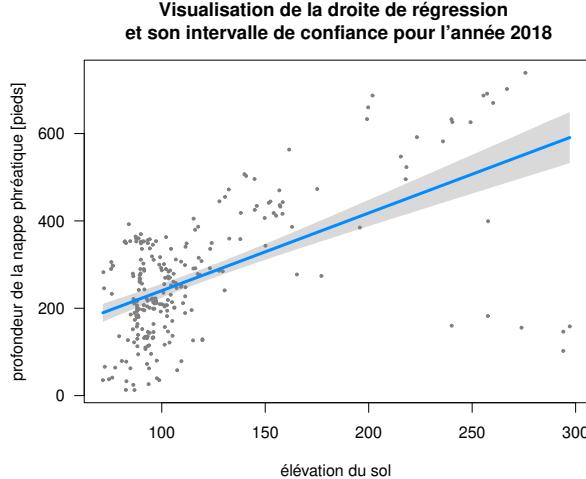
A.1 Droites de régression linéaires



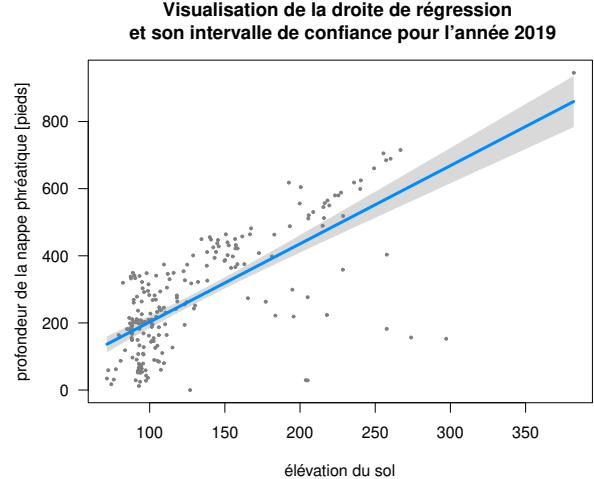
(a) 2016



(b) 2017



(c) 2018



(d) 2019

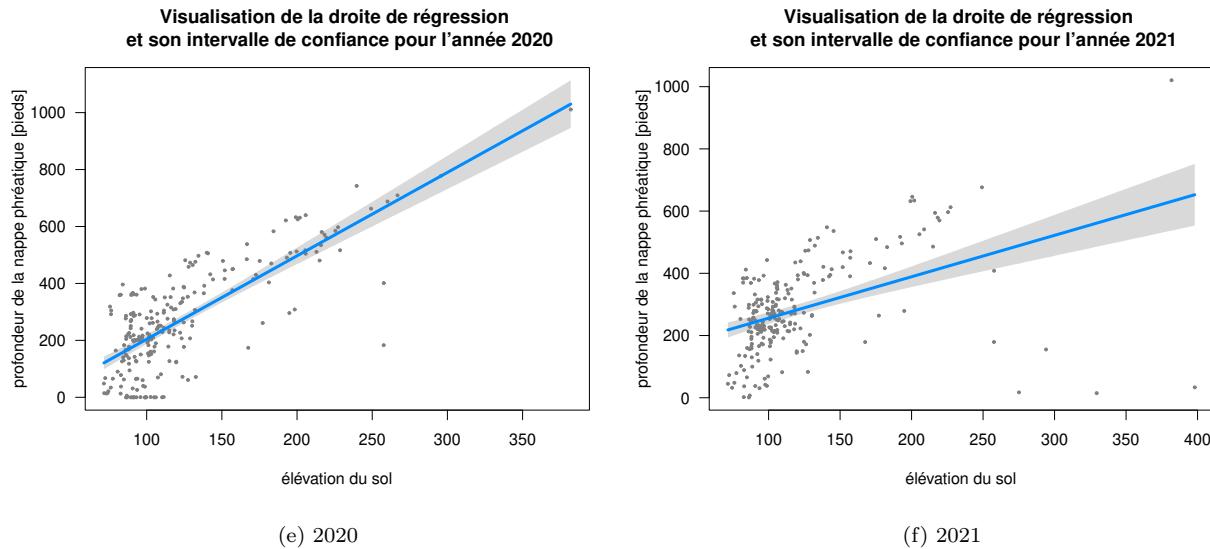
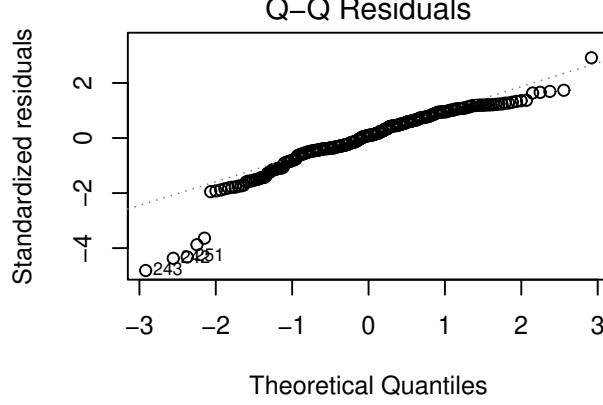
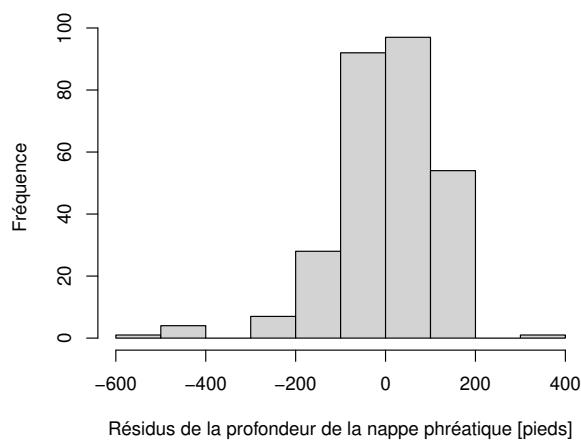
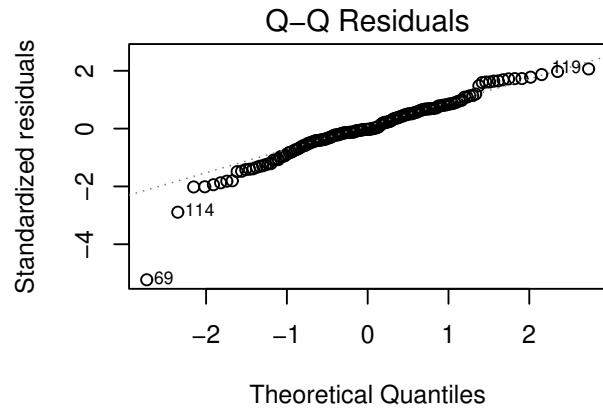
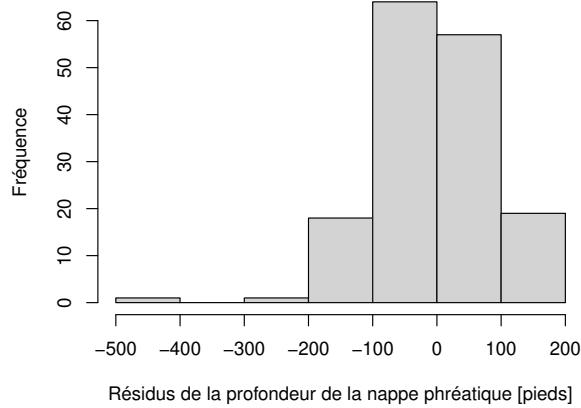


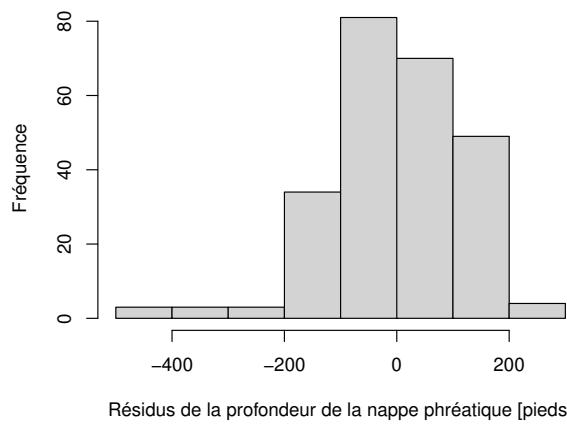
FIGURE 11 – Droites de régression linéaires de la profondeur de la nappe phréatique en fonction de l’élévation du sol à Bakersfield pour tous les mois d’octobre entre 2016 et 2021

TABLE 5 – Estimateurs des paramètres des droites de régression linéaire pour chaque année entre 2016 et 2021

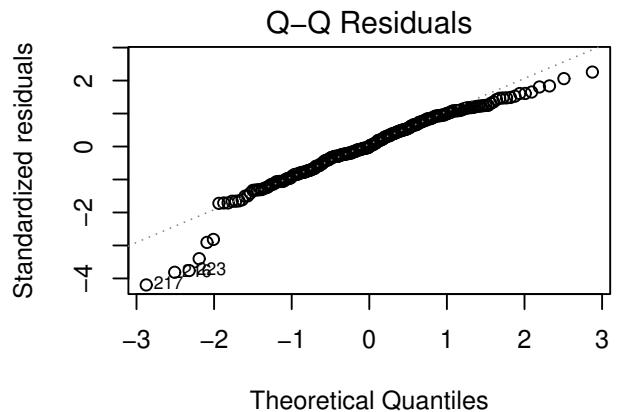
Année	$\hat{\beta}_0$	$\hat{\beta}_1$
2016	49.989	2.728
2017	38.4552	2.0166
2018	62.4284	1.7781
2019	-30.2223	2.3294
2020	-88.7983	2.9287
2021	122.675	1.331

A.2 Distribution des résidus

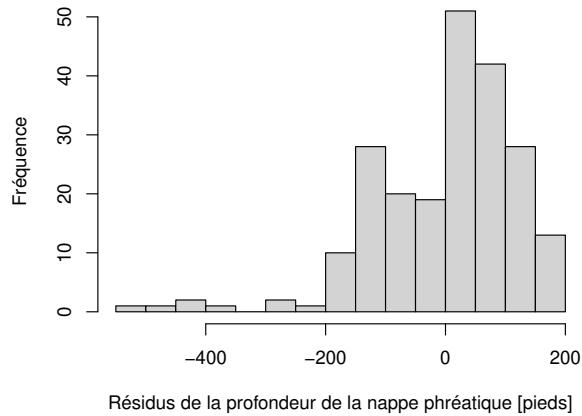




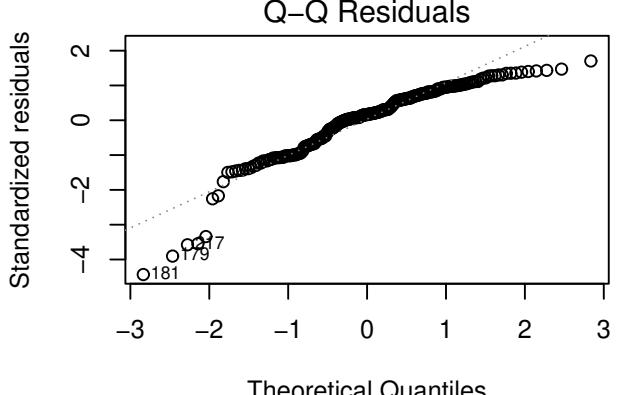
(e) 2018



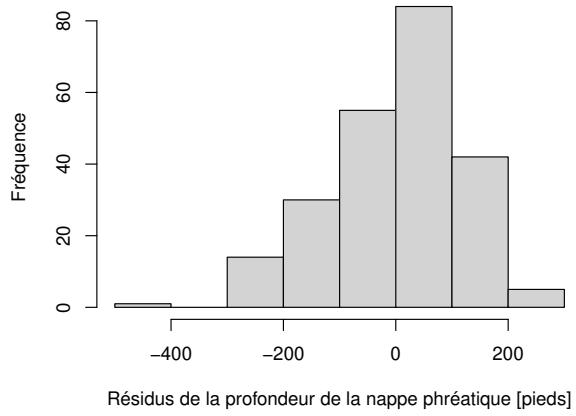
(f) 2018



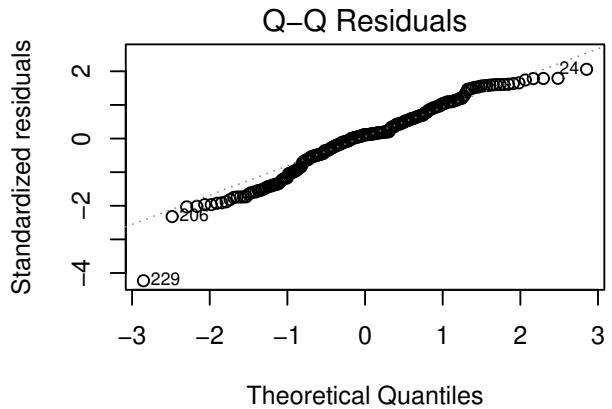
(g) 2019



(h) 2019

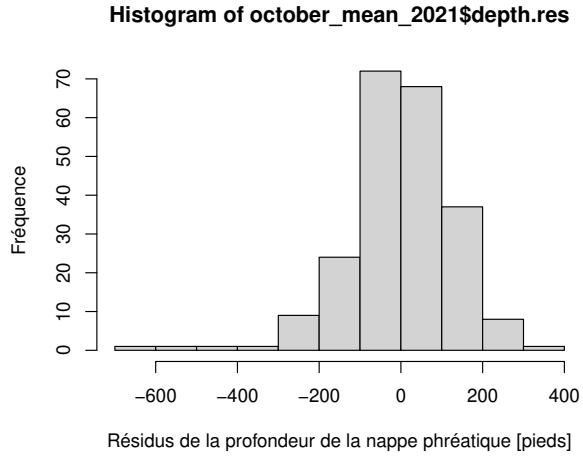


(i) 2020

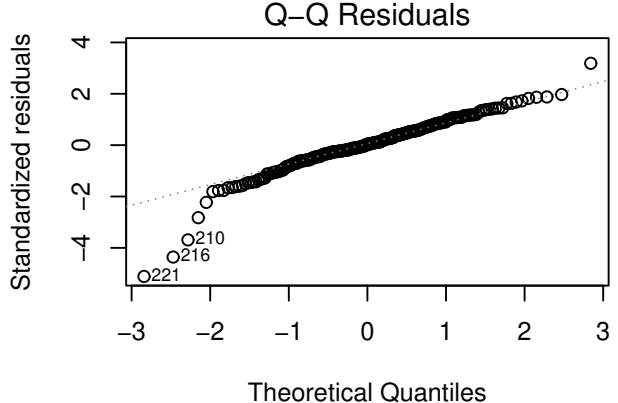


Theoretical Quantiles

(j) 2020



(k) 2021



Theoretical Quantiles

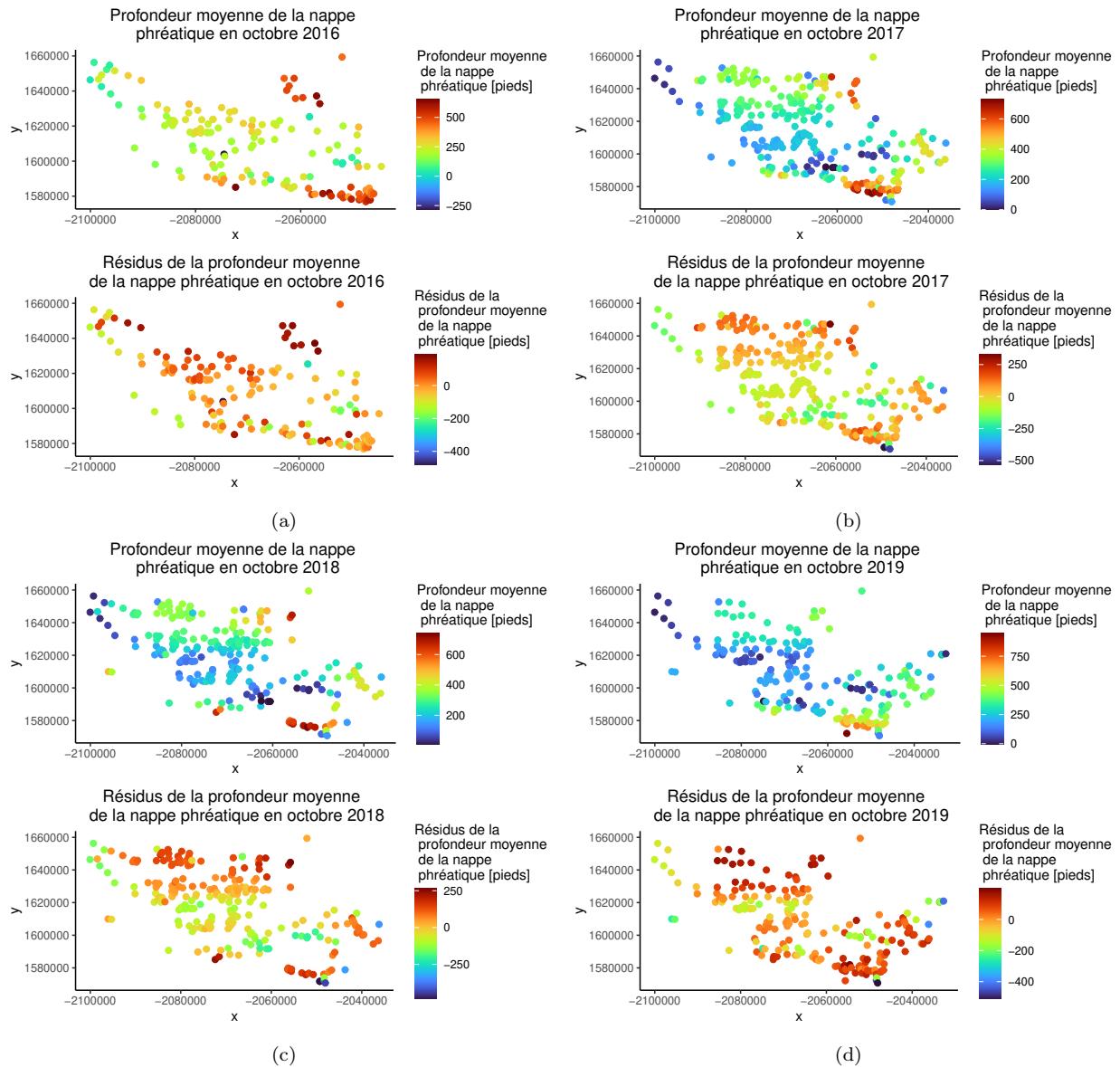
(l) 2021

FIGURE 12 – Histogrammes des résidus obtenus après avoir retiré l'effet de l'élévation du sol sur la profondeur de la nappe phréatique à Bakersfield en octobre 2016 à 2021 (figures a, c, e, g, i et k) et Q-Q plots de ces résidus (figures b, d, f, h, j et l)

TABLE 6 – Tableau reprenant les coefficients de détermination linéaire de Pearson (R^2) associés à chaque modèle de régression linéaire entre 2015 et 2021

Année	2016	2017	2018	2019	2020	2021
R²	61.7 %	43.12 %	33.8 1%	52.5 %	59.13 %	20.2 %

A.3 Répartition spatiale de la profondeur de la nappe avant et après avoir retiré l'effet de l'altitude du sol



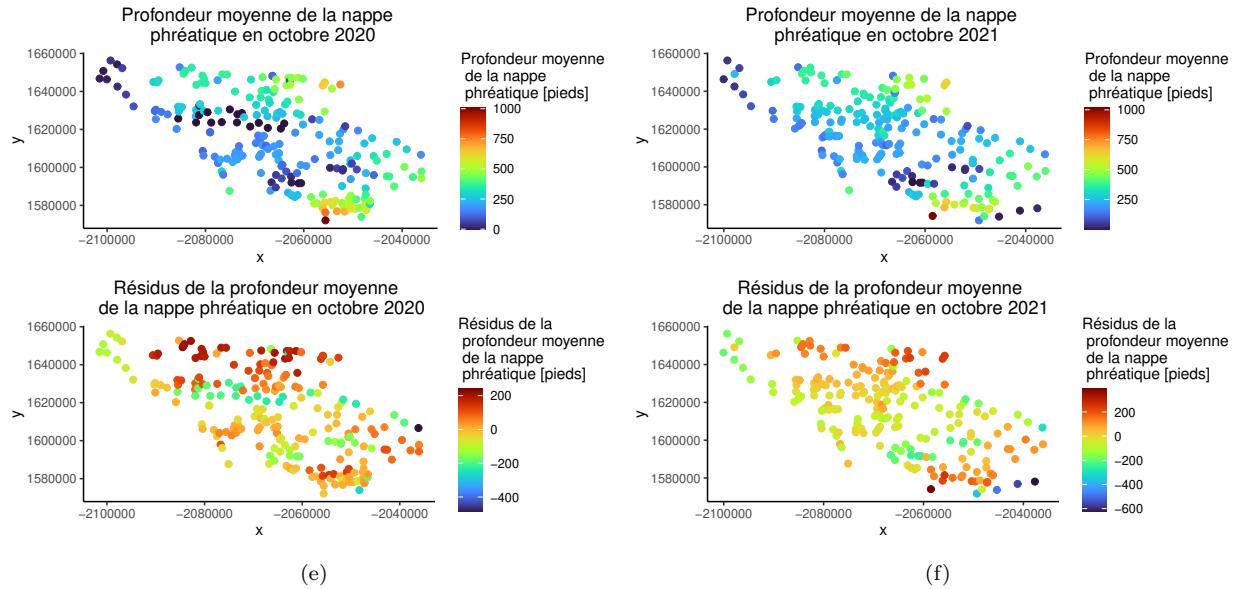


FIGURE 13 – Carte de la profondeur (en haut) et des résidus de la profondeur (en bas) moyenne de la nappe phréatique en octobre entre 2016 et 2021 à toute les stations de mesure autour de Bakersfield.

B Variogrammes

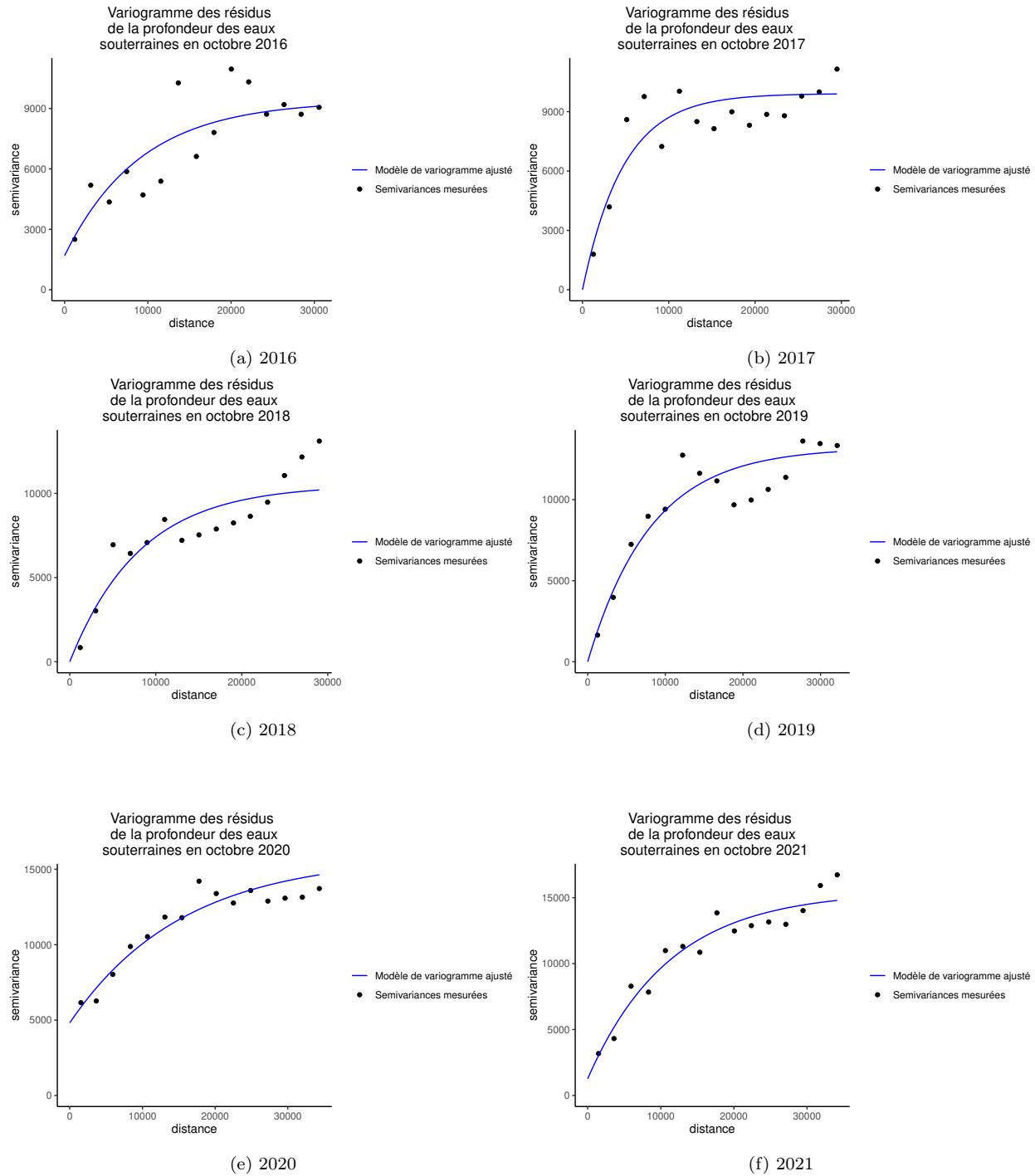


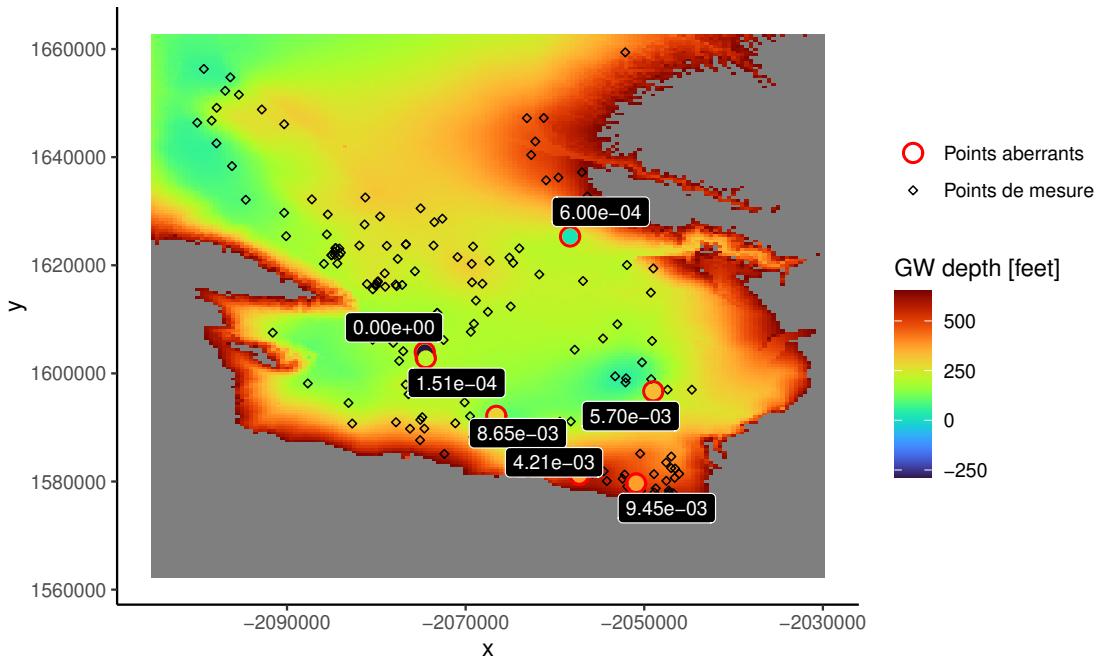
FIGURE 14 – Variogrammes expérimentaux et modélisés pour toutes les années entre 2016 et 2021

TABLE 7 – Valeurs optimisées des paramètres du modèle de variogramme obtenu à partir des données de profondeur de la nappe en octobre 2016 à 2021

Année	Modèle	Palier ($\hat{\sigma}^2$)	Portée ($\hat{\theta}$)
2016	Pépite	1690.209	0
	Exponentiel	7694.074	9116.381
2017	Pépite	0	0
	Exponentiel	9914.799	4746.093
2018	Pépite	0	0
	Exponentiel	10504.4	8182.205
2019	Pépite	0	0
	Exponentiel	13211.67	8148.11
2020	Pépite	4816.535	0
	Exponentiel	11012.069	15468.83
2021	Pépite	1270.675	0
	Exponentiel	14196.910	11194.5

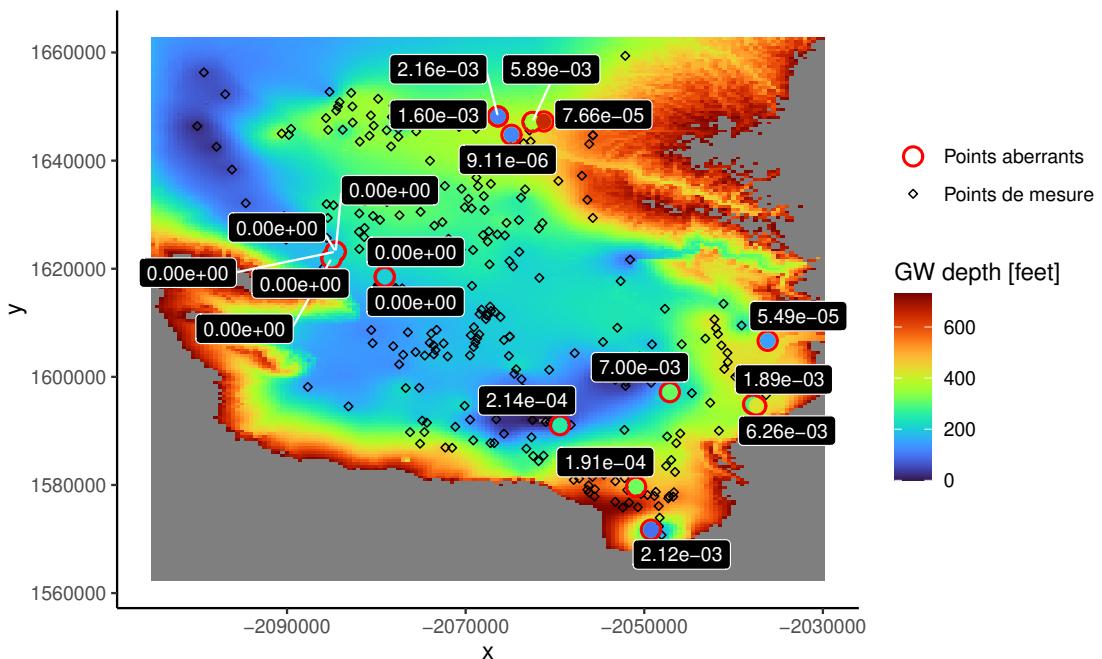
C Cartes de prédition obtenues par krigeage

Prédictions par krigeage de la profondeur de la nappe phréatique en octobre 2016 et points aberrants avec leur p–valeurs associées



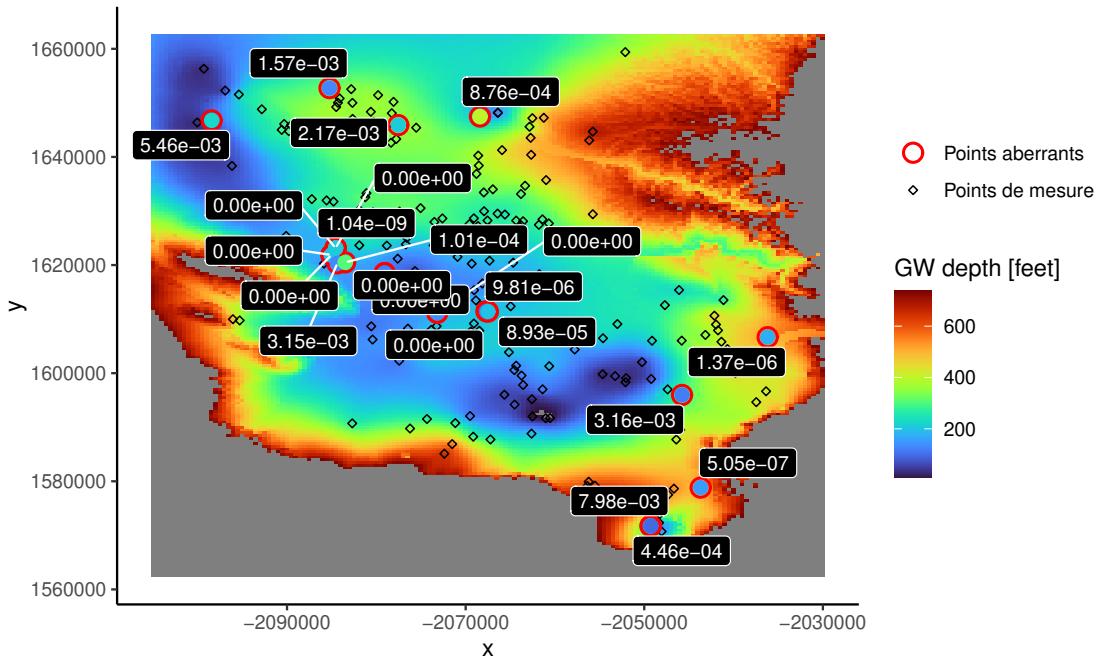
(a) 2016

Prédictions par krigeage de la profondeur de la nappe phréatique en octobre 2017 et points aberrants avec leur p–valeurs associées



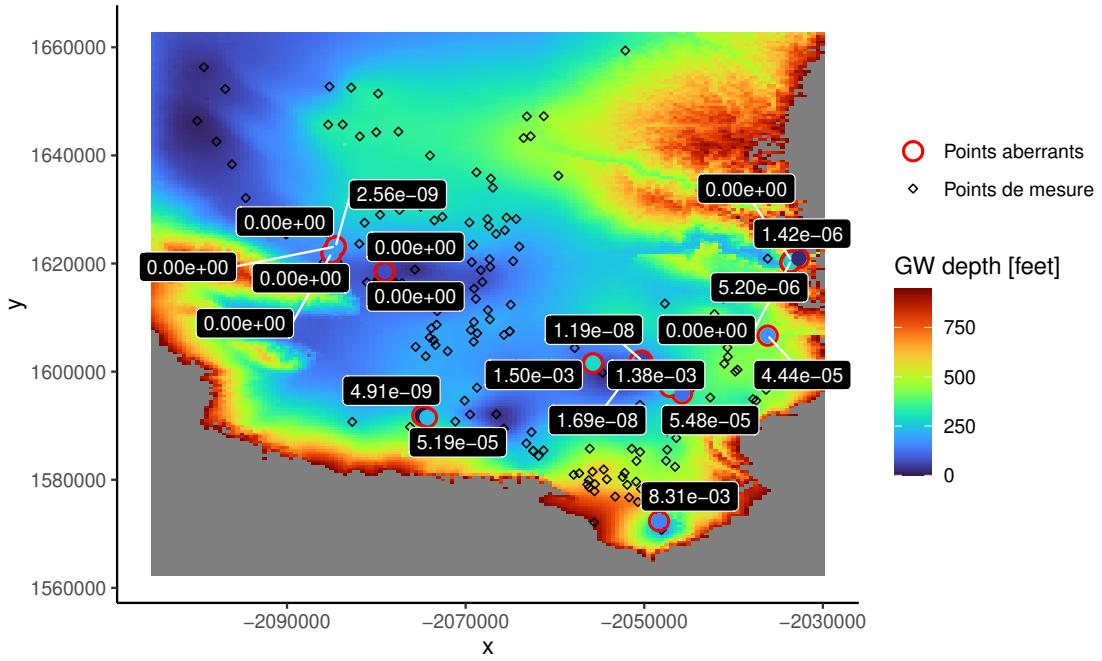
(b) 2017

Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2018 et points aberrants avec leur p–valeurs associées



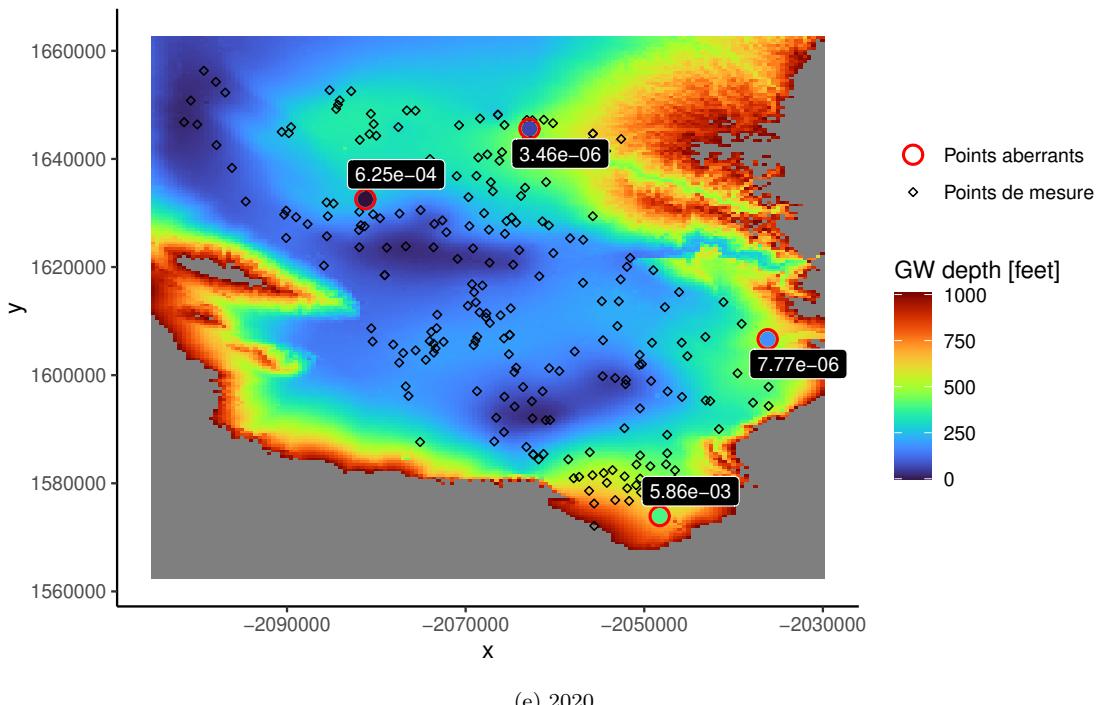
(c) 2018

Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2019 et points aberrants avec leur p–valeurs associées



(d) 2019

Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2020 et points aberrants avec leur p–valeurs associées



Prédictions par krigage de la profondeur de la nappe phréatique en octobre 2021 et points aberrants avec leur p–valeurs associées

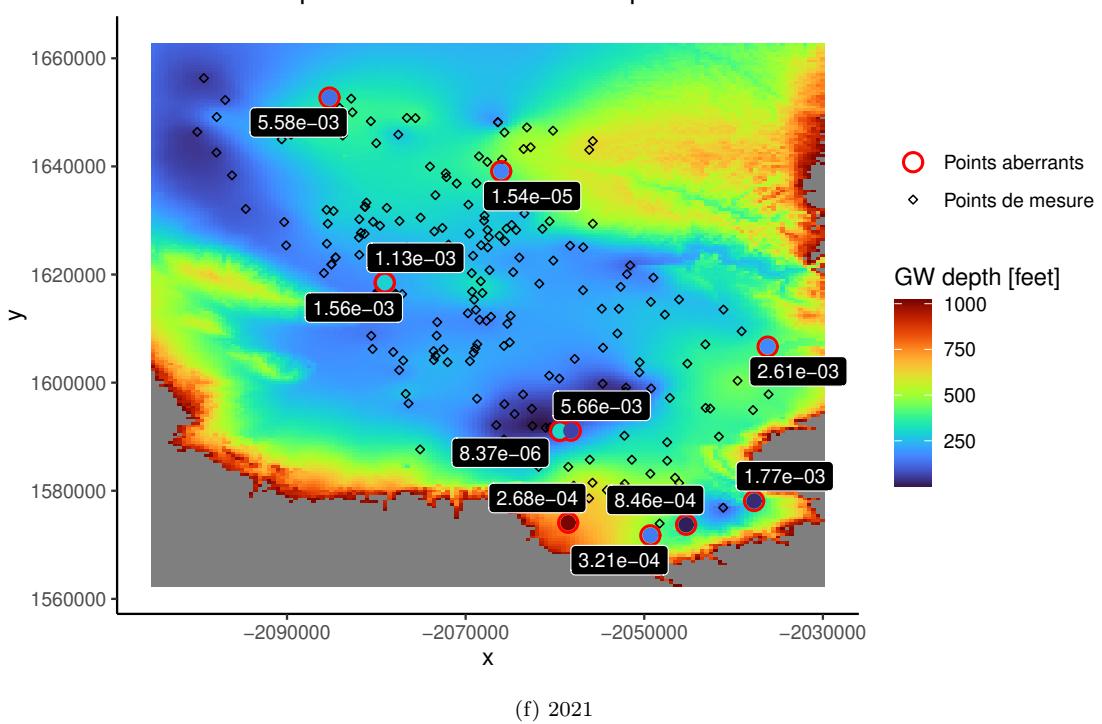


FIGURE 15 – Cartes des prédictions de la profondeur de la nappe pour chaque mois d'octobre entre 2016 et 2021.

D Programmes réalisés pour les différentes analyses

L'entièreté du code a été écrit en R (version 4.4.1). Le script utilisé pour réaliser toutes les analyses et créer toutes les figures de ce rapport se trouvent ci-dessous. ChatGPT et Copilot ont été utilisées pour debugger certaines parties du code, notamment la série temporelle et la carte des tendances.

```
# list of packages used for running the code chunks below
library(ggplot2)
library(ggnewscale)
library(ggrepel)
library(gridExtra)
library(car)
library(MASS)
library(PerformanceAnalytics)
library(stats)
library(tidyverse)
library(viridis)
library(lubridate)
library(data.table)
library(fields)
library(gstat)
library(pander)
library(visreg)
library(data.table)
library(matrixStats)

theme_set(theme_classic())

setwd("~/Documents/Data science in bioscience engineering/
Projet_California")

# Loading the data
monthly_mean_delimited <- read.csv(file =
"monthly_mean_delimited.csv", header = TRUE, sep = ",",
row.names = 1)
DEM_grid <- read.csv(file = "DEM_grid.csv", header = TRUE, sep =
", ")

head(monthly_mean_delimited)
str(monthly_mean_delimited)

# Creating a dataframe that contains all the duplicates (same year,
# same month, same coordinates (x, y) but different station names and
# mean_gse_gwe values)
```

```

duplicates <- monthly_mean_delimited %>%
  group_by(x, y, year, month) %>%
  filter(n() > 1) %>%
  ungroup()

# Viewing the duplicates
duplicates %>%
  filter(year == 2015 & month == 10) %>%
  as.data.frame() %>%
  head()

# Jeu de données original mais avec les réplicats remplacés par une
seule ligne dont la mean_gse_gwe est la moyenne de tous les
réplicats.

# Le premier site_code dans chaque groupe de réplicats est considéré
comme le site_code de la nouvelle ligne
combined_data <- monthly_mean_delimited %>%
  group_by(x, y, year, month) %>%
  summarize(
    mean_gse_gwe = mean(mean_gse_gwe, na.rm = TRUE),
    site_code = first(site_code),
    longitude = first(longitude),
    latitude = first(latitude),
    .groups = 'drop'
  )

# Rearrangement de l'ordre des colonnes
combined_data <- combined_data %>%
  dplyr::select(site_code, year, month, mean_gse_gwe, x, y,
longitude, latitude)
# Le nouveau dataframe n'a plus de réplicats
combined_data %>%
  group_by(x, y, year, month) %>%
  filter(n() > 1) %>%
  ungroup() %>%
  filter(year == 2015 & month == 10) %>%
  as.data.frame() %>%
head()

# Duplicated stations:
# exemple: stations 356025N1192413W001 et 356025N1192413W002 à 1cm
l'une de l'autre, mais même latitude et longitude... (en octobre
2015)

```

```

monthly_mean_delimited %>%
  filter(year == 2015 & month == 10) %>%
  filter(site_code == "356025N1192413W001" | site_code ==
"356025N1192413W002") %>%
  head()

# Mais on décide de n'enlever que les duplcats pour x et y (pas
pour la longitude et latitude) car ce ont fait krigage sur
coordonnées x et y et c'est là que les duplcats posent problème
# Remplacement du jeu de données de départ par le jeu de données
sans les réplcats

monthly_mean_delimited <- combined_data
# On passe d'un jeu de données avec 102459 lignes à un jeu de
données avec 95715 lignes.

# Valeurs mesurées durant les mois d'octobre entre 2015 et 2021

october_mean <- monthly_mean_delimited[monthly_mean_delimited$month
== 10 & monthly_mean_delimited$year %in% 2015:2021,]

summary(october_mean)

# Grouping the data into the different stations and calculating the
mean depth for each station
mean_depth <- october_mean %>%
  group_by(site_code, x, y, longitude, latitude) %>%
  summarise(site_mean_depth = mean(mean_gse_gwe))

# Converting the tibble object into a data frame object
mean_depth <- as.data.frame(mean_depth)
head(mean_depth)

# Selecting the station 353539N1191118W001
selected_station = mean_depth[mean_depth$site_code ==
"353539N1191118W001",]

# Dessin de la carte des profondeurs moyennes + altitude + encercler
point de station
ggplot(mean_depth) +
  geom_tile(data = DEM_grid, aes(x = x, y = y, fill = elevation)) +
  scale_fill_gradient(name = "Elévation du sol [pieds]", low =
"white", high = "black") +
  geom_point(aes(x = x, y = y, color = site_mean_depth), size = 2) +
  scale_color_viridis(name = "Profondeur moyenne \n de la nappe \n"

```

```

phréatique [pieds]", option = 'turbo') +
  geom_point(data = selected_station, aes(x, y),
             pch = 21, fill = NA, size = 4, colour = "red",
             stroke = 1) +
  labs(x = "x" , y = "y" , title = "Carte de la profondeur moyenne
de la nappe phréatique en octobre \n entre 2015 et 2021 à toutes les
stations de mesure \n et élévation du sol autour de Bakersfield") +
  theme(plot.title = element_text(hjust = 0.5))

# Série temporelle profondeur nappe
station_monthly_mean <-
monthly_mean_delimited[monthly_mean_delimited$site_code ==
"353539N1191118W001",]
station_monthly_mean$date <-
as.Date(paste(station_monthly_mean$year, station_monthly_mean$month,
"01", sep="-"), "%Y-%m-%d")

head(station_monthly_mean)

# Plotting of the time series
ggplot(station_monthly_mean) + geom_point(aes(x = date, y =
mean_gse_gwe)) +
  scale_x_date(date_breaks = "2 years", date_labels = "%Y") +
  labs(x = "Date", y = "Profondeur moyenne mensuelle \n de la nappe
phréatique [pieds]") +
  scale_y_reverse( ) +
  ggtitle("Niveau de la nappe phréatique pour la station
353539N1191118W001") +
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x =
element_text(angle = 90, hjust = 0.5, vjust = 0.5))

# Créer un fonction qui fait la régression linéaire pour n'importe
quelle année entre 2015 et 2021:
linear_reg <- function(year){
  # Selecting the right year in the data frame
  yearly_october_mean <- october_mean[october_mean$year == year,]

  # Creating a grid for the elevation prediction (at the measurement
  stations)
  stations_grid <- yearly_october_mean[c("x","y")]

  # Predicting the elevation at the stations and adding it to the
  data frame
}

```

```

elevation_voronoi <- idw(formula = elevation ~ 1, data = DEM_grid,
                           locations = ~ x + y, newdata = stations_grid,
                           nmax = 1)

yearly_october_mean$elevation <- elevation_voronoi$var1.pred

# Doing the linear regression
mod <- lm(mean_gse_gwe ~ elevation, data = yearly_october_mean)

# Printing the results
print(summary(mod))
print("Confidence intervals")
print(confint(mod))
print(anova(mod))

# Checking if the linear regression hypotheses are met
# (normal distribution of the residues, homoscedasticity)
par(mfrow=c(2,2))
plot(mod)

# Visualising the linear regression
par(mfrow=c(1,1))
visreg(mod, main = paste("Visualisation de la droite de régression
\n et son intervalle de confiance pour l'année", year), xlab =
"élévation du sol", ylab = "profondeur de la nappe phréatique
[pieds]")

# Calculating the predicted depth from the linear model
xpred <- data.frame(elevation = yearly_october_mean$elevation)
pred <- predict(mod, xpred, interval = "prediction")
yearly_october_mean$depth.pred <- pred
yearly_october_mean$depth.res <- residuals(mod)
lm_results <- list(yearly_october_mean, mod)
return(lm_results)
}

# Fonction qui dessine la répartition spatiale de la profondeur de
la nappe avant et après avoir retiré l'effet de l'altitude du sol

residuals_plot <- function(yearly_october_mean){
  year <- yearly_october_mean$year[1]

  plot_mean_depth <- ggplot(yearly_october_mean) +
    geom_point(aes(x = x, y = y, color = mean_gse_gwe), size = 2) +
    scale_color_viridis(name = "Profondeur moyenne \n de la nappe \n"

```

```

phréatique [pieds]", option = 'turbo') +
  labs(x = "x" , y = "y" , title = paste("Profondeur moyenne de la
nappe \n phréatique en octobre", year)) +
  theme(plot.title = element_text(hjust = 0.5))

plot_res <- ggplot(yearly_october_mean) +
  geom_point(aes(x = x, y = y, color = depth.res), size = 2) +
  scale_color_viridis(name = "Résidus de la \n profondeur
moyenne \n de la nappe \n phréatique [pieds]", option = 'turbo') +
  labs(x = "x" , y = "y" , title = paste("Résidus de la
profondeur moyenne \n de la nappe phréatique en octobre", year)) +
  theme(plot.title = element_text(hjust = 0.5))
grid.arrange(plot_mean_depth, plot_res)
}

# Linear regression for the year 2015
lm_results_2015 <- linear_reg(2015)
october_mean_2015 <- lm_results_2015[[1]]
lm_2015 <- lm_results_2015[[2]]
head(october_mean_2015)
ggplot(october_mean_2015) +
  geom_point(aes(x = elevation, y = depth.res)) +
  geom_hline(aes(yintercept = 0))
hist(october_mean_2015$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")
ggplot(october_mean_2015) +
  geom_histogram(aes(depth.res), bins = 11, color = "black", fill =
"darkorange")

# Linear regression for the year 2021
lm_results_2021 <- linear_reg(2021)
october_mean_2021 <- lm_results_2021[[1]]
lm_2021 <- lm_results_2021[[2]]

head(october_mean_2021)

ggplot(october_mean_2021) +
  geom_point(aes(x = elevation, y = depth.res)) +
  geom_hline(aes(yintercept = 0))

hist(october_mean_2021$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")

# Régression linéaire pour les autres années:
lm_results_2016 <- linear_reg(2016)
october_mean_2016 <- lm_results_2016[[1]]

```

```

lm_2016 <- lm_results_2016[[2]]
hist(october_mean_2016$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")

lm_results_2017 <- linear_reg(2017)
october_mean_2017 <- lm_results_2017[[1]]
lm_2017 <- lm_results_2017[[2]]
hist(october_mean_2017$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")

lm_results_2018 <- linear_reg(2018)
october_mean_2018 <- lm_results_2018[[1]]
lm_2018 <- lm_results_2018[[2]]
hist(october_mean_2018$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")

lm_results_2019 <- linear_reg(2019)
october_mean_2019 <- lm_results_2019[[1]]
lm_2019 <- lm_results_2019[[2]]
hist(october_mean_2019$depth.res, breaks = 20, xlab = "Résidus de la
profondeur de la nappe phréatique [pieds]", ylab = "Fréquence")

lm_results_2020 <- linear_reg(2020)
october_mean_2020 <- lm_results_2020[[1]]
lm_2020 <- lm_results_2020[[2]]

hist(october_mean_2020$depth.res, xlab = "Résidus de la profondeur
de la nappe phréatique [pieds]", ylab = "Fréquence")

residuals_plot(october_mean_2015)
residuals_plot(october_mean_2016)
residuals_plot(october_mean_2017)
residuals_plot(october_mean_2018)
residuals_plot(october_mean_2019)
residuals_plot(october_mean_2020)
residuals_plot(october_mean_2021)

# Variogramme expérimental - fonction qui dessine le variogramme
# expérimental des résidus
# et retourne l'objet res.vario (avec les données du variogramme
# expérimental).

exp_variogram <- function(yearly_october_mean, hmax){

year <- yearly_october_mean$year[1]

```

```

res.gstat <- gstat(formula = depth.res ~ 1, data =
yearly_october_mean, locations = ~x+y)

res.vario <- variogram(res.gstat, cutoff = hmax)
variance_res <- var(yearly_october_mean$depth.res)

# Plot the experimental variogram
vario_plot <- ggplot(res.vario) +
  geom_point(aes(x = dist, y = gamma, color = "Semivariances
mesurées")) +
  geom_hline(aes(yintercept = variance_res, linetype = "Variance
des résidus")) +
  labs(title = paste("Variogramme expérimental des \n résidus de
la profondeur des eaux \n souterraines en octobre", year, sep =
"), x = "distance", y = "semivariance") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits = c(0, max(res.vario$dist))) +
  scale_y_continuous(limits = c(0, max(res.vario$gamma,
variance_res))) +
  scale_color_manual(values = c("Semivariances mesurées" =
"black", "Fitted variogram model" = "blue"), name = NULL) +
  scale_linetype_manual(values = c("Variance des résidus" =
"dashed"), name = NULL)

print(vario_plot)

return(res.vario)
}

# Fonction qui calcule les paramètres du variogramme
# Les paramètres du modèle de variogramme ont des valeurs par défaut
# qui peuvent être changées
model_variogram <- function(res.vario, year, psill = 8000, model =
"Exp", range = 35000, nugget = 2400){

  res.vario.model <- vgm(psill, model, range, nugget)
  res.fit.model <- fit.variogram(res.vario, model = res.vario.model)
  res.vario.model.values <- variogramLine(res.fit.model, maxdist =
max(res.vario$dist))

  model_vario_plot <- ggplot(res.vario) +
    geom_point(aes(x = dist, y = gamma, color = "Semivariances
mesurées")) +
    geom_line(data = res.vario.model.values, aes(x = dist, y =
gamma, color = "Modèle de variogramme ajusté")) +

```

```

    labs(title = paste("Variogramme des résidus \n de la profondeur
des eaux \n souterraines en octobre", year, sep = " "), x =
"distance", y = "semivariance") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits = c(0, max(res.vario$dist))) +
  scale_y_continuous(limits = c(0, max(res.vario$gamma,
res.vario.model.values$gamma))) +
  scale_color_manual(values = c("Semivariances mesurées" =
"black", "Modèle de variogramme ajusté" = "blue"), name = NULL)
  print(model_vario_plot)

  return(res.fit.model)
}

# Variogramme expérimental et modélisé pour octobre 2015
hmax_2015 <- max(dist(october_mean_2015[, c("x", "y")]))

res.vario_2015 <- exp_variogram(october_mean_2015, hmax_2015)
res.fit_model_2015 <- model_variogram(res.vario_2015, 2015, psill =
8000, range = 35000, nugget = 2400, model = "Exp")

print(res.fit_model_2015)

# Variogramme expérimental et modélisé pour octobre 2021
hmax_2021 <- max(dist(october_mean_2021[, c("x", "y")]))

res.vario_2021 <- exp_variogram(october_mean_2021)
res.fit_model_2021 <- model_variogram(res.vario_2021, 2021, psill =
18000, range = 45000, nugget = 2400, model = "Exp")

print(res.fit_model_2021)

# Création d'une fonction qui trouve les valeurs extraordinaire
dans le jeu de données (niveau 1 - alpha = 0.99 pour l'intervalle de
prédition)

# La fonction retourne une liste: le premier élément est un vecteur
contenant les indices des valeurs extraordinaire dans le jeu de
données yearly_october_mean (ex: october_mean_2015).

# Le 2e élément de la liste contient toutes les p-valeurs

find_outliers <- function(yearly_october_mean, res.fit_model){
  outliers <- c()
  pvalue <- c()
  alpha <- 0.01
}

```

```

Z <- qnorm(1-alpha/2)
for(i in 1:nrow(yearly_october_mean)){
  pred.i <- krige(formula = depth.res ~ 1, # Kriging on the
residuals
    data = yearly_october_mean[-i,],
    locations = ~x+y,
    newdata = yearly_october_mean[i,c("x","y")],
    model = res.fit_model)
  pred.interval <- c(pred.i$var1.pred - Z*sqrt(pred.i$var1.var),
pred.i$var1.pred + Z*sqrt(pred.i$var1.var))
  outliers[i] <- ifelse(yearly_october_mean$depth.res[i] <
pred.interval[1] | yearly_october_mean$depth.res[i] >
pred.interval[2],1,0)
  zscore <- (yearly_october_mean$depth.res[i] - pred.i$var1.pred)/
sqrt(pred.i$var1.var)
  pvalue[i] <- 2 * (1 - pnorm(abs(zscore)))
}
which_out <- which(outliers == 1)

results <- list(which_out, pvalue)

return(results)
}

# Loading the prediction grid
depth.grid <- read.csv(file = "grid.csv", header = TRUE, sep = ",")
depth.grid$X <- NULL # remove 1st column

# Fonction qui trace la carte des prédictions de la profondeur de la
nappe pour l'année souhaitée
# Les valeurs extraordinaires précédemment identifiées sont marquées
en rouge sur la carte
# La fonction retourne les résultats du krigeage (sans les outliers)
sous forme d'un dataframe

plot_krig <- function(yearly_october_mean, res.fit_model, lm_year,
which_out, pvalue){

  # res_fit_model est le modèle du variogramme des résidus

  # Krigeage (pour les résidus) sans les valeurs aberrantes
  res.krig.out <- krige(formula = depth.res~1,
    data = yearly_october_mean[-which_out,],
    loc = ~x+y,

```

```

    newdata = depth.grid,
    model = res.fit_model)

# Résultats du krigage pour la profondeur de la nappe phréatique
# (prédicteur = résidu + valeur de la fonction espérance)

# La grille d'interpolation depth.grid contient les mêmes points
(x, y) que DEM_grid
xpred <- data.frame(elevation = DEM_grid$elevation)
depth.pred <- predict(lm_year, xpred, interval = "prediction")
depth.pred <- as.data.frame(depth.pred)

# On obtient le prédicteur en additionnant le résultat du krigage
pour les résidus aux valeurs d'espérance de la profondeur
res.krig.out$depth.pred <- depth.pred$fit + res.krig.out$var1.pred

min_depth <- min(yearly_october_mean$mean_gse_gwe)

max_depth <- max(yearly_october_mean$mean_gse_gwe)

p <- ggplot() +
  geom_tile(data = res.krig.out, aes(x = x, y = y, fill =
depth.pred)) +
  geom_point(data = yearly_october_mean[-which_out,],
             aes(x = x, y = y, color = "Points de mesure"),
             shape = 5, size = 1) +
  geom_point(data = yearly_october_mean[which_out,],
             aes(x, y, color = "Points aberrants"),
             pch = 21, fill = NA, size = 3.5, stroke = 1) +
  scale_color_manual(values = c("Points aberrants" = "red",
                                "Points de mesure" = "black"),
name = NULL) +
  new_scale_color() +
  geom_point(data = yearly_october_mean[which_out,],
             aes(x, y, color = mean_gse_gwe),
             size = 2, stroke = 1, show.legend = FALSE) +
  scale_color_viridis(name = NULL, option = 'turbo', limits =
c(min_depth, max_depth)) +
  geom_label_repel(data = yearly_october_mean[which_out, ],
                   aes(x = x, y = y,
                       label = sprintf("%.2e", pvalue[which_out])),
                   size = 3, color = "white", fill = "black",
box.padding = unit(0.25, "lines"),
nudge_x = 0.05, nudge_y = -0.1,
max.overlaps = getOption("ggrepel.max.overlaps"),

```

```

default = 20)) +
  scale_fill_viridis(name = "Profondeur \n de la nappe \n
phréatique [pieds]", option = 'turbo',
                      limits = c(min_depth, max_depth)) +
  labs(title = paste("Prédictions par krigeage de la profondeur de
la nappe phréatique en \n octobre",
                     yearly_october_mean$year[1], "et points
aberrants avec leur p-valeurs associées", sep = " "),
       x = "x",
       y = "y") +
  theme(plot.title = element_text(hjust = 0.5))

print(p)

  return(res.krig.out)
}

# Valeurs extraordinaires / aberrantes pour l'année 2015:
outliers_2015 <- find_outliers(october_mean_2015,
                                 res.fit_model_2015)
(which_out_2015 <- outliers_2015[[1]])
(pvalue_2015 <- outliers_2015[[2]])

krig_2015 <- plot_krig(october_mean_2015, res.fit_model_2015,
                        lm_2015, which_out_2015, pvalue_2015)
paste("Expected number of outliers found with this method :
", nrow(october_mean_2015)*0.01)
paste("Number of outliers found with the LOOCV method:",
      length(which_out_2015))

# 2016
# Max distance to estimate the variogram
hmax_2016 <- max(dist(october_mean_2016[, c("x", "y")])) / 3
res.vario_2016 <- exp_variogram(october_mean_2016, hmax_2016)

res.fit_model_2016 <- model_variogram(res.vario_2016, 2016, psill =
8000, range = 20000, nugget = 1000, model = "Exp")
print(res.fit_model_2016)

outliers_2016 <- find_outliers(october_mean_2016,
                                 res.fit_model_2016)
(which_out_2016 <- outliers_2016[[1]])
(pvalue_2016 <- outliers_2016[[2]])
krig_2016 <- plot_krig(october_mean_2016, res.fit_model_2016,
                        lm_2016, which_out_2016, pvalue_2016)

```

```

# Station 352015N1192094W001 --> profondeur = -285.7 ????

paste("Expected number of outliers found with this method :",
",nrow(october_mean_2016)*0.01)
paste("Number of outliers found with the LOOCV method:", 
length(which_out_2016))

# 2017
hmax_2017 <- max(dist(october_mean_2017[, c("x", "y")])) / 3
res.vario_2017 <- exp_variogram(october_mean_2017, 30500)

res.fit_model_2017 <- model_variogram(res.vario_2017, 2017, psill =
9000, range = 30000, nugget = 0, model = "Exp")

print(res.fit_model_2017)

outliers_2017 <- find_outliers(october_mean_2017,
res.fit_model_2017)
(which_out_2017 <- outliers_2017[[1]])
(pvalue_2017 <- outliers_2017[[2]])
krig_2017 <- plot_krig(october_mean_2017, res.fit_model_2017,
lm_2017, which_out_2017, pvalue_2017)

paste("Expected number of outliers found with this method :",
",nrow(october_mean_2017)*0.01)
paste("Number of outliers found with the LOOCV method:", 
length(which_out_2017))

# 2018
hmax_2018 <- max(dist(october_mean_2018[, c("x", "y")])) / 3
res.vario_2018 <- exp_variogram(october_mean_2018, 30000)

res.fit_model_2018 <- model_variogram(res.vario_2018, 2018, psill =
15000, range = 35000, nugget = 0, model = "Exp")

print(res.fit_model_2018)
outliers_2018 <- find_outliers(october_mean_2018,
res.fit_model_2018)
(which_out_2018 <- outliers_2018[[1]])
(pvalue_2018 <- outliers_2018[[2]])
krig_2018 <- plot_krig(october_mean_2018, res.fit_model_2018,
lm_2018, which_out_2018, pvalue_2018)

```

```

paste("Expected number of outliers found with this method : 
",nrow(october_mean_2018)*0.01)
paste("Number of outliers found with the LOOCV method:", 
length(which_out_2018))

# 2019
hmax_2019 <- max(dist(october_mean_2019[, c("x", "y")])) / 3
res.vario_2019 <- exp_variogram(october_mean_2019, hmax_2019)
res.fit_model_2019 <- model_variogram(res.vario_2019, 2019, psill = 
8000, range = 35000, nugget = 2400, model = "Exp")
print(res.fit_model_2019)
outliers_2019 <- find_outliers(october_mean_2019,
res.fit_model_2019)

(which_out_2019 <- outliers_2019[[1]])
(pvalue_2019 <- outliers_2019[[2]])
krig_2019 <- plot_krig(october_mean_2019, res.fit_model_2019,
lm_2019, which_out_2019, pvalue_2019)

paste("Expected number of outliers found with this method : 
",nrow(october_mean_2019)*0.01)

paste("Number of outliers found with the LOOCV method:", 
length(which_out_2019))

# 2020
res.vario_2020 <- exp_variogram(october_mean_2020)
res.fit_model_2020 <- model_variogram(res.vario_2020, 2020, psill = 
8000, range = 35000, nugget = 2400, model = "Exp")
print(res.fit_model_2020)
outliers_2020 <- find_outliers(october_mean_2020,
res.fit_model_2020)

(which_out_2020 <- outliers_2020[[1]])
(pvalue_2020 <- outliers_2020[[2]])
krig_2020 <- plot_krig(october_mean_2020, res.fit_model_2020,
lm_2020, which_out_2020, pvalue_2020)

paste("Expected number of outliers found with this method : 
",nrow(october_mean_2020)*0.01)
paste("Number of outliers found with the LOOCV method:", 
length(which_out_2020))

# 2021:
outliers_2021 <- find_outliers(october_mean_2021,

```

```

res.fit_model_2021)

(which_out_2021 <- outliers_2021[[1]])
(pvalue_2021 <- outliers_2021[[2]])

krig_2021 <- plot_krig(october_mean_2021, res.fit_model_2021,
lm_2021, which_out_2021, pvalue_2021)

paste("Expected number of outliers found with this method :
",nrow(october_mean_2021)*0.01)
paste("Number of outliers found with the LOOCV method:",
length(which_out_2021))

# Outlier stations for each year
print(october_mean_2015[which_out_2015, "site_code"])

# Create a list of all flagged indices for each year
outlier_lists <- list(
  "2015" = which_out_2015,
  "2016" = which_out_2016,
  "2017" = which_out_2017,
  "2018" = which_out_2018,
  "2019" = which_out_2019,
  "2020" = which_out_2020,
  "2021" = which_out_2021
)

# Initialize an empty list to store outlier data for each year
outlier_data <- list()

# Loop through each year's outlier indices
for (year in names(outlier_lists)) {
  # Get the indices of outliers for the current year
  indices <- outlier_lists[[year]]

  # Dynamically refer to the corresponding dataset for the current
  year
  dataset_name <- paste0("october_mean_", year)
  current_dataset <- get(dataset_name)

  # Extract site code, X, and Y for these indices from the
  respective October mean dataset

  year_outliers <- current_dataset[indices, c("site_code", "x",
  "y", "mean_gse_gwe")]

```

```

# Add a column to specify the year
year_outliers$year <- as.integer(year)
# Append to the list
outlier_data[[year]] <- year_outliers
}

# Combine all years into a single dataset
outlier_dataset <- do.call(rbind, outlier_data)

# Access the dataset for a specific year (e.g., 2015)
october_mean_2015_outliers <- outlier_data[["2015"]]
repeating_stations <- october_mean_2015_outliers %>%
  group_by(site_code) %>%
  summarize(
    occurrences = n(),
    x = first(x),
    y = first(y)
  ) %>%
  filter(occurrences > 1)

repeated_stations <- outlier_dataset %>%
  group_by(site_code) %>%
  summarize(repeat_count = n_distinct(year)) %>%
  filter(repeat_count > 1) %>%
  pull(site_code)
outlier_dataset <- outlier_dataset %>%
  mutate(is_repeated = ifelse(site_code %in% repeated_stations,
TRUE, FALSE))

# Count occurrences of each station
flagged_station_counts <- outlier_dataset %>%
  group_by(site_code, x, y) %>%
  summarize(flag_count = n(), .groups = "drop")

# Merge counts back into the dataset for visualization
outlier_dataset <- outlier_dataset %>%
  left_join(flagged_station_counts, by = c("site_code", "x", "y"))
ggplot() +
  # DEM background
  geom_tile(data = DEM_grid, aes(x = x, y = y, fill = elevation)) +
  scale_fill_gradient(name = "Élévation du sol [pieds]", low =
"white", high = "black") +
  # Plot outlier points, sized by flag count and colored by flag
  # count (gradient: green -> red)

```

```

geom_point(data = outlier_dataset, aes(x = x, y = y, color =
flag_count), size = 4) +
  scale_color_gradient2(
    name = "Nombre de signalements",
    low = "green3", mid = "yellow", high = "red", midpoint = 3 # 
Transition from green (low flag counts) to red (high flag counts)
)+

# Customize labels and title
  labs(
    x = "x",
    y = "y",
    title = "Carte des stations aberrantes de 2015 à 2021 \n avec
nombre de signalements" )+
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "right", # Move legends to the right
    legend.box = "vertical" # Arrange legends vertically
  )+ guides(
    color = guide_legend(
      override.aes = list(size = 5), # Make legend points for flag
      count larger
      order = 1, title.position = "top", title.hjust = 0.5
    ),
    size = guide_legend(
      order = 2, title.position = "top", title.hjust = 0.5
    ),
    fill = guide_colorbar(
      order = 3, title.position = "top", title.hjust = 0.5
    )
  )

# Variance de prédition pour chaque année / chaque mois d'octobre

prediction_variance <- cbind(krig_2015[,c("x", "y")],
                                krig_2015$var1.var,
                                krig_2016$var1.var,
                                krig_2017$var1.var,
                                krig_2018$var1.var,
                                krig_2019$var1.var,
                                krig_2020$var1.var,
                                krig_2021$var1.var)
colnames(prediction_variance) <- c("x", "y",
as.character(2015:2021))

```

```

mean_pred_var <- rowMeans(prediction_variance[, -c(1, 2)])
prediction_variance <- cbind(prediction_variance, average =
mean_pred_var)

ggplot() +
  geom_tile(data = prediction_variance, aes(x = x, y = y, fill =
average)) +
  geom_point(data = october_mean,
             aes(x = x, y = y, color = "Points de mesure"), pch =
3, size = 1, alpha = 0.5) +
  scale_fill_viridis(name = "Variance de prédition [pieds^2]",
option = 'turbo') +
  scale_color_manual(values = c("Points de mesure" = "black"),
name = NULL) +
  labs(title = paste("Moyenne des variances de prédition des
résidus
de la profondeur de la nappe pour les 7 mois
d'octobre entre 2015 et 2021"),
x = "x",
y = "y") +
  theme(plot.title = element_text(hjust = 0.5))

# lgd_name <- expression(paste("Variance de prédition",
\text{pieds}^{\{2\}}))

krig_all_years <- cbind(DEM_grid[, 1:2],
                         krig_2015$depth.pred,
                         krig_2016$depth.pred,
                         krig_2017$depth.pred,
                         krig_2018$depth.pred,
                         krig_2019$depth.pred,
                         krig_2020$depth.pred,
                         krig_2021$depth.pred)
colnames(krig_all_years) <- c("x", "y", as.character(2015:2021))

lm_data <- cbind(year = 2015:2021, t(krig_all_years[,-c(1,2)]))
# Initialisation d'un data frame vide pour stocker les valeurs de la
pente et de la p-valeur pour chaque pixel

depth_trends <- data.frame(slope = rep(NA, ncol(lm_data) - 1),
pvalue = rep(NA, ncol(lm_data) - 1))

for (i in 1:(ncol(lm_data) - 1)){
  pixel <- as.data.frame(lm_data[, c(1, i + 1)])

```

```

colnames(pixel) <- c("year", "depth")
res <- lm(depth ~ year, data = pixel)
depth_trends[i, ] <- coef(summary(res))[2,c("Estimate", "Pr(>|t|)")]
}
depth_trends <- cbind(DEM_grid[,c("x", "y")], depth_trends)

significant_stations <- depth_trends %>%
  filter(pvalue < 0.05)

which_significant <- which(depth_trends$pvalue <= 0.05)

ggplot(depth_trends) +
  geom_tile(aes(x = x, y = y, fill = slope)) +
  scale_fill_gradientn(
    name = "Tendances (2015-2021)",
    colors = c("blue", "white", "pink", "red"),
    values = scales::rescale(c(min(depth_trends$slope), 0, 10,
max(depth_trends$slope))) )+
  labs(title = "Tendance de la profondeur de la nappe phréatique
\n en octobre entre 2015 et 2021",
x = "x",
y = "y") +
  geom_point(data = significant_stations,
             aes(x = x, y = y, color = "tendances \n significatives
\n (p-valeur < 0.05)"),
             pch = 3, size = 0.75) +
  scale_color_manual(values = c("tendances \n significatives \n (p-
valeur < 0.05)" = "black"), name = NULL) +
  theme(plot.title = element_text(hjust = 0.5))

#colors = c("darkblue", "blue", "deepskyblue", "lightblue", "white",
#"mistyrose", "lightcoral", "indianred", "firebrick", "red"),
#values = scales::rescale(c(min(depth_trends$slope), -200, -150,
-100, 0, 5, 10, 15, 18, max(depth_trends$slope)))

# Même graphe mais avec limits sur l'échelle
ggplot(depth_trends) +
  geom_tile(aes(x = x, y = y, fill = slope)) +
  scale_fill_gradientn(
    name = "Tendances (2015-2021)",
    colors = c("blue", "white", "red"),
    values = scales::rescale(c(-50, 0, max(depth_trends$slope))),
    limits = c(-50, max(depth_trends$slope)))
)+
```

```
labs(title = "Tendance de la profondeur de la nappe phréatique  
\n en octobre entre 2015 et 2021",  
      x = "x",  
      y = "y") +  
  geom_point(data = significant_stations, aes(x = x, y = y, color =  
"tendances \n significatives \n (p-valeur < 0.05)"), pch = 3, size =  
1) +  
  scale_color_manual(values = c("tendances \n significatives \n (p-  
valeur < 0.05)" = "black"), name = NULL) +  
  theme(plot.title = element_text(hjust = 0.5))
```