****

**MTH8302**

**Modèles de régression et d’analyse de la variance**

**Devoir4 - Examen Final**

**distribution : dimanche 24 juin 2018 - 9h00**

**remise : lundi 02 juillet 2018 - 12h00 (au plus tard)**

**Ce travail est réalisé individuellement par chaque étudiant inscrit au cours.**

**Chaque étudiant le fait SEUL sans demander de l’aide à d’autres.**

**En apposant sa signature sur son rapport soumis, l’étudiant (e) certifie sur son honneur avoir fait ce travail seul. L’obtention des résultats présentés et la rédaction de ce travail ne fait l’objet d’aucun plagiat, partiel ou total.**

**Information concernant le plagiat à Polytechnique :** [**http://www.polymtl.ca/etudes/ppp/index.php**](http://www.polymtl.ca/etudes/ppp/index.php)

**Exigences pour la rédaction du rapport consulter la page 4 du plan de cours**

[**http://www.groupes.polymtl.ca/mth6301/mth8302/2018-MTH8302-PlanCours.pdf**](http://www.groupes.polymtl.ca/mth6301/mth8302/2018-MTH8302-PlanCours.pdf)

**Transmettez cette page comme la page 1 de votre rapport; une copie est disponible sur**

[**http://www.groupes.polymtl.ca/mth6301/MTH8301-babillard.htm**](http://www.groupes.polymtl.ca/mth6301/MTH8301-babillard.htm)

**Compléter l’information suivante.**

**MTH8302 Modèles de régression et d’analyse de variance**

**NOM \_\_BETTACHE\_\_\_\_\_ PRÉNOM \_Lyes Heythem\_**

**MATRICULE \_** **1923715\_ SIGNATURE \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Transmettre votre rapport par courriel à** [**bernard.clement@polymtl.ca**](file:///C:\Users\Bernard\Documents\07.MTH8302\0.Devoirs&Exer\2014\bernard.clement@polymtl.ca)

**CORRECTION**

|  |  |  |
| --- | --- | --- |
| **QUESTION** | **POIDS** | **OBTENU** |
| **B** | **40** |  |
| **C** | **30** |  |
| **D** | **10** |  |
| **E** | **10** |  |
| **qualité** | **10** |  |
| **TOTAL** | **100** |  |
| **Bonus** | **10** |  |

**Le devoir4 est décrit dans les pages suivantes. Il s’agit d’une étude de cas comme convenu,**

Table des matières

[**Résumé :** 6](#_Toc518331696)

[A) Création des 2 groupes 7](#_Toc518331697)

[B) Développement des modèles sur l’ensemble train 8](#_Toc518331698)

[Var40 :Y\_GPER\_BDO 8](#_Toc518331699)

[RÉGRESSION MULTIPLE 8](#_Toc518331700)

[la méthode de Régression avec Backward Stepwise 11](#_Toc518331701)

[la méthode de Régression avec Backward Stepwise sans les points aberrants 13](#_Toc518331702)

[Modèle de Régression MARS 15](#_Toc518331703)

[Réseaux de neurones (SANN 18](#_Toc518331704)

[VAR41 : Y\_GPERF\_CDO 20](#_Toc518331705)

[RÉGRESSION MULTIPLE 20](#_Toc518331706)

[la méthode de Régression avec Backward Stepwise 23](#_Toc518331707)

[Modèle de Régression MARS 25](#_Toc518331708)

[Réseaux de neurones (SANN) 28](#_Toc518331709)

[VAR42 : Y\_GPERF\_SS 30](#_Toc518331710)

[RÉGRESSION MULTIPLE 30](#_Toc518331711)

[la méthode de Régression avec Backward Stepwise 33](#_Toc518331712)

[la méthode de Régression avec Backward Stepwise sans les points aberrants 35](#_Toc518331713)

[Modèle de Régression MARS 37](#_Toc518331714)

[C) Comparer la performance de nos modèles sur l’ensemble train et sur l’ensemble test. 42](#_Toc518331715)

[Var40 :Y\_GPER\_BDO sur l’ensemble test 42](#_Toc518331716)

[la méthode de Régression avec Backward Stepwise 42](#_Toc518331717)

[Modèle de Régression MARS 44](#_Toc518331718)

[Réseaux de neurones (SANN) 46](#_Toc518331719)

[VAR41 : Y\_GPERF\_CDO sur l’ensemble test 48](#_Toc518331720)

[la méthode de Régression avec Backward Stepwise 48](#_Toc518331721)

[Modèle de Régression MARS 50](#_Toc518331722)

[Réseaux de neurones (SANN) 52](#_Toc518331723)

[VAR42 : Y\_GPERF\_SS sur l’ensemble test 54](#_Toc518331724)

[la méthode de Régression avec Backward Stepwise 54](#_Toc518331725)

[Modèle de Régression MARS 56](#_Toc518331726)

[Réseaux de neurones (SANN ) 58](#_Toc518331727)

[Comparer la performance de nos modèles sur l’ensemble train : 60](#_Toc518331728)

[Var40 :Y\_GPER\_BDO 60](#_Toc518331729)

[Var41 :Y\_GPER\_CDO 61](#_Toc518331730)

[Var42 :Y\_GPER\_SS 62](#_Toc518331731)

[Comparer la performance de vos modèles sur l’ensemble test 63](#_Toc518331732)

[Var40 :Y\_GPER\_BDO 63](#_Toc518331733)

[Var41 :Y\_GPER\_CDO 63](#_Toc518331734)

[Var42 :Y\_GPER\_SS 64](#_Toc518331735)

[D) une conclusion sommaire 65](#_Toc518331736)

[E) Comparer les données d’entrée de 1990 et celles de 1991 66](#_Toc518331737)

[Conlusion générale : 69](#_Toc518331738)

**Liste des graphiques**

[Figure 1: Pareto Chart of t-Values for Coefficients 8](#_Toc518308801)

[Figure 2: les valeurs Prédites avec les valeurs Observes 9](#_Toc518308802)

[Figure 3: Analyse des résidus avec les valeurs Observes 9](#_Toc518308803)

[Figure 4: résidus sur échelle de probabilité gaussienne 10](#_Toc518308804)

[Figure 5:Analyse des résidus avec les valeurs Prédites 10](#_Toc518308805)

[Figure 6: Pareto Chart of t-Values for Coefficients 11](#_Toc518308806)

[Figure 7:les valeurs Prédites avec les valeurs Observes 11](#_Toc518308807)

[Figure 8: résidus sur échelle de probabilité gaussienne 12](#_Toc518308808)

[Figure 9: Analyse des résidus avec les valeurs Observes 12](#_Toc518308809)

[Figure 10: Analyse des résidus avec les valeurs Prédites 12](#_Toc518308810)

[Figure 11: Pareto Chart of t-Values for Coefficients 13](#_Toc518308811)

[Figure 12: les valeurs Prédites avec les valeurs Observes 14](#_Toc518308812)

[Figure 13: Analyse des résidus avec les valeurs Observes 14](#_Toc518308813)

[Figure 14:résidus sur échelle de probabilité gaussienne 14](#_Toc518308814)

[Figure 15: Analyse des résidus avec les valeurs Prédites 14](#_Toc518308815)

[Figure 16: : les valeurs Prédites avec les valeurs Observes 16](#_Toc518308816)

[Figure 17: : Analyse des résidus avec les valeurs Prédites 16](#_Toc518308817)

[Figure 18: Y\_GPER\_BDO avec FL 17](#_Toc518308818)

[Figure 19: Y\_GPER\_BDO avec BDO 17](#_Toc518308819)

[Figure 20: Y\_GPER\_BDO avec BDO\_S 17](#_Toc518308820)

[Figure 21: Y\_GPER\_BDO avec CDO\_S 17](#_Toc518308821)

[Figure 22:Y\_GPER\_BDO avec Y\_PER\_SED\_P 17](#_Toc518308822)

[Figure 23: Y\_GPER\_BDO avec Y\_PER\_BDO\_S 17](#_Toc518308823)

[Figure 24: Y\_GPER\_BDO avec Y\_PER\_CDO\_S 18](#_Toc518308824)

[Figure 25:les valeurs Prédites modèle 1 avec les valeurs Observes 18](#_Toc518308825)

[Figure 26: les valeurs Prédites modèle 2 avec les valeurs Observes 18](#_Toc518308826)

[Figure 27: les valeurs Prédites avec les valeurs Observes 19](#_Toc518308827)

[Figure 28: Analyse des résidus avec les valeurs Prédites modèle 1 19](#_Toc518308828)

[Figure 29: Analyse des résidus avec les valeurs Prédites modèle 2 19](#_Toc518308829)

[Figure 30: Pareto Chart of t-Values for Coefficients 20](#_Toc518308830)

[Figure 31: les valeurs Prédites avec les valeurs Observes 21](#_Toc518308831)

[Figure 32: Analyse des résidus avec les valeurs Observes 21](#_Toc518308832)

[Figure 33: résidus sur échelle de probabilité gaussienne 22](#_Toc518308833)

[Figure 34:Analyse des résidus avec les valeurs Prédites 22](#_Toc518308834)

[Figure 35: Pareto Chart of t-Values for Coefficients 23](#_Toc518308835)

[Figure 36: les valeurs Prédites avec les valeurs Observes 23](#_Toc518308836)

[Figure 37:résidus sur échelle de probabilité gaussienne 24](#_Toc518308837)

[Figure 38: Analyse des résidus avec les valeurs Prédites 24](#_Toc518308838)

[Figure 39: les valeurs Prédites avec les valeurs Observes 26](#_Toc518308839)

[Figure 40: Analyse des résidus avec les valeurs Prédites 26](#_Toc518308840)

[Figure 41: Y\_GPER\_CDO avec CDO 27](#_Toc518308841)

[Figure 42: Y\_GPER\_CDO avec BDO\_S 27](#_Toc518308842)

[Figure 43: Y\_GPER\_CDO avec CDO\_S 27](#_Toc518308843)

[Figure 44: Y\_GPER\_CDO avec Y\_PER\_SS\_P 27](#_Toc518308844)

[Figure 45: Y\_GPER\_CDO avec Y\_PER\_SED\_P 27](#_Toc518308845)

[Figure 46: Y\_GPER\_CDO avec Y\_PER\_BDO\_S 27](#_Toc518308846)

[Figure 47: Y\_GPER\_CDO avec Y\_PER\_CDO\_S 28](#_Toc518308847)

[Figure 48: les valeurs Prédites modèle 1 avec les valeurs Observes 28](#_Toc518308848)

[Figure 49: les valeurs Prédites modèle 2 avec les valeurs Observes 28](#_Toc518308849)

[Figure 50: les valeurs Prédites avec les valeurs Observes 29](#_Toc518308850)

[Figure 51: Analyse des résidus avec les valeurs Prédites modèle 1 29](#_Toc518308851)

[Figure 52: Analyse des résidus avec les valeurs Prédites modèle 2 29](#_Toc518308852)

[Figure 53: Pareto Chart of t-Values for Coefficients 30](#_Toc518308853)

[Figure 54: les valeurs Prédites avec les valeurs Observes 30](#_Toc518308854)

[Figure 55: Analyse des résidus avec les valeurs Observes 31](#_Toc518308855)

[Figure 56: résidus sur échelle de probabilité gaussienne 31](#_Toc518308856)

[Figure 57: Analyse des résidus avec les valeurs Prédites 32](#_Toc518308857)

[Figure 58 :les valeurs Prédites avec les valeurs Observes 33](#_Toc518308858)

[Figure 59: résidus sur échelle de probabilité gaussienne 34](#_Toc518308859)

[Figure 60: Analyse des résidus avec les valeurs Observes 34](#_Toc518308860)

[Figure 61: Analyse des résidus avec les valeurs Prédites 34](#_Toc518308861)

[Figure 62: Analyse des résidus avec les valeurs Observes 35](#_Toc518308862)

[Figure 63: les valeurs Prédites avec les valeurs Observes 36](#_Toc518308863)

[Figure 64: Analyse des résidus avec les valeurs Prédites 36](#_Toc518308864)

[Figure 65: résidus sur échelle de probabilité gaussienne 36](#_Toc518308865)

[Figure 66: les valeurs Prédites avec les valeurs Observes 38](#_Toc518308866)

[Figure 67: Analyse des résidus avec les valeurs Prédites 38](#_Toc518308867)

[Figure 68: Y\_GPER\_SS avec SS 39](#_Toc518308868)

[Figure 69: Y\_GPER\_SS avec COND\_P 39](#_Toc518308869)

[Figure 70: Y\_GPER\_SS avec BDO\_S 39](#_Toc518308870)

[Figure 71: Y\_GPER\_SS avec Y\_PER\_SS\_P 39](#_Toc518308871)

[Figure 72: Y\_GPER\_SS avec Y\_PER\_BDD\_S 39](#_Toc518308872)

[Figure 73: Y\_GPER\_SS avec Y\_PER\_CDO\_S 39](#_Toc518308873)

[Figure 74: es valeurs Prédites modèle 1 avec les valeurs Observes 40](#_Toc518308874)

[Figure 75:les valeurs Prédites modèle 2 avec les valeurs Observes 40](#_Toc518308875)

[Figure 76: les valeurs Prédites avec les valeurs Observes 41](#_Toc518308876)

[Figure 77 :Analyse des résidus avec les valeurs Prédites modèle 1 41](#_Toc518308877)

[Figure 78:Analyse des résidus avec les valeurs Prédites modèle 2 41](#_Toc518308878)

[Figure 79: Pareto Chart of t-Values for Coefficients 42](#_Toc518308879)

[Figure 80: les valeurs Prédites avec les valeurs Observes 42](#_Toc518308880)

[Figure 81:résidus sur échelle de probabilité gaussienne 43](#_Toc518308881)

[Figure 82: Analyse des résidus avec les valeurs Observes 43](#_Toc518308882)

[Figure 83: Analyse des résidus avec les valeurs Prédites 43](#_Toc518308883)

[Figure 84:les valeurs Prédites avec les valeurs Observes 45](#_Toc518308884)

[Figure 85:Analyse des résidus avec les valeurs Prédites 45](#_Toc518308885)

[Figure 86:les valeurs Prédites modèle 1 avec les valeurs Observes 46](#_Toc518308886)

[Figure 87: les valeurs Prédites modèle 2 avec les valeurs Observes 46](#_Toc518308887)

[Figure 88:Pareto Chart of t-Values for Coefficients 48](#_Toc518308888)

[Figure 89:les valeurs Prédites avec les valeurs Observes 48](#_Toc518308889)

[Figure 90:résidus sur échelle de probabilité gaussienne 49](#_Toc518308890)

[Figure 91: Analyse des résidus avec les valeurs Observes 49](#_Toc518308891)

[Figure 92:Analyse des résidus avec les valeurs Prédites 49](#_Toc518308892)

[Figure 93: les valeurs Prédites avec les valeurs Observes 51](#_Toc518308893)

[Figure 94: Analyse des résidus avec les valeurs Prédites 51](#_Toc518308894)

[Figure 95:les valeurs Prédites modèle 1 avec les valeurs Observes 52](#_Toc518308895)

[Figure 96: les valeurs Prédites modèle 2 avec les valeurs Observes 52](#_Toc518308896)

[Figure 97:les valeurs Prédites avec les valeurs Observes 53](#_Toc518308897)

[Figure 98:Analyse des résidus avec les valeurs Prédites modèle 1 53](#_Toc518308898)

[Figure 99:Analyse des résidus avec les valeurs Prédites modèle 2 53](#_Toc518308899)

[Figure 100:Pareto Chart of t-Values for Coefficients 54](#_Toc518308900)

[Figure 101:les valeurs Prédites avec les valeurs Observes 54](#_Toc518308901)

[Figure 102: résidus sur échelle de probabilité gaussienne 55](#_Toc518308902)

[Figure 103:Analyse des résidus avec les valeurs Observes 55](#_Toc518308903)

[Figure 104: Analyse des résidus avec les valeurs Prédites 55](#_Toc518308904)

[Figure 105:les valeurs Prédites avec les valeurs Observes 57](#_Toc518308905)

[Figure 106: Analyse des résidus avec les valeurs Prédites 57](#_Toc518308906)

[Figure 107:les valeurs Prédites modèle 1 avec les valeurs Observes 58](#_Toc518308907)

[Figure 108: les valeurs Prédites modèle 2 avec les valeurs Observes 58](#_Toc518308908)

[Figure 109:les valeurs Prédites avec les valeurs Observes 59](#_Toc518308909)

[Figure 110:Analyse des résidus avec les valeurs Prédites modèle 1 59](#_Toc518308910)

[Figure 111: Analyse des résidus avec les valeurs Prédites modèle 2 59](#_Toc518308911)

[Figure 112: Box & Whisker Plot: pH 67](#_Toc518308912)

[Figure 113: Box & Whisker Plot: SS 67](#_Toc518308913)

[Figure 114: Box & Whisker Plot: SED 67](#_Toc518308914)

[Figure 115: Box & Whisker Plot: VSS 67](#_Toc518308915)

[Figure 116: Box & Whisker Plot: COND 67](#_Toc518308916)

[Figure 117: Box & Whisker Plot: BDO 67](#_Toc518308917)

[Figure 118: Box & Whisker Plot: CDO 68](#_Toc518308918)

[Figure 119: Box & Whisker Plot: ZN 68](#_Toc518308919)

[Figure 120: Box & Whisker Plot: FL 68](#_Toc518308920)

**Liste des tableaux**

[**Tableau 1:les 10 premières valeurs obtenues de aléa** 7](#_Toc518308921)

[**Tableau 2: ID de l'ensemble test** 7](#_Toc518308922)

[Tableau 3:Tests d'importance univariés pour Y\_GPER\_BDO 8](#_Toc518308923)

[Tableau 4 : Paramètre Estimâtes et VIF 9](#_Toc518308924)

[Tableau 5: R,Radj, SS Model et SS residual 9](#_Toc518308925)

[Tableau 6: Tests d'importance univariés pour Y\_GPER\_BDO 11](#_Toc518308926)

[Tableau 7: Paramètre Estimâtes et VIF 11](#_Toc518308927)

[Tableau 8: R,Radj, SS Model et SS residual 11](#_Toc518308928)

[Tableau 9:Paramètre Estimâtes et VIF 13](#_Toc518308929)

[Tableau 10: R,Radj, SS Model et SS residual 13](#_Toc518308930)

[Tableau 11: Number of times each predictor is referenced 15](#_Toc518308931)

[Tableau 12: Regression statistics 15](#_Toc518308932)

[Tableau 13: Model coefficients 16](#_Toc518308933)

[Tableau 14: Summary of active networks 18](#_Toc518308934)

[Tableau 15: Tests d'importance univariés pour Y\_GPER\_CDO 20](#_Toc518308935)

[Tableau 16 : Paramètre Estimâtes et VIF 21](#_Toc518308936)

[Tableau 17: R,Radj, SS Model et SS residual 21](#_Toc518308937)

[Tableau 18: Tests d'importance univariés pour Y\_GPER\_CDO 23](#_Toc518308938)

[Tableau 19: Paramètre Estimâtes et VIF 23](#_Toc518308939)

[Tableau 20: R,Radj, SS Model et SS residual 23](#_Toc518308940)

[Tableau 21: Number of times each predictor is referenced 25](#_Toc518308941)

[Tableau 22: Regression statistics 25](#_Toc518308942)

[Tableau 23: Model coefficients 26](#_Toc518308943)

[Tableau 24: Summary of active networks 28](#_Toc518308944)

[Tableau 25: Tests d'importance univariés pour Y\_GPERF\_SS 30](#_Toc518308945)

[Tableau 26: Paramètre Estimâtes et VIF 30](#_Toc518308946)

[Tableau 27: R,Radj, SS Model et SS residual 30](#_Toc518308947)

[Tableau 28: Tests d'importance univariés pour Y\_GPEF\_SS 33](#_Toc518308948)

[Tableau 29: Paramètre Estimâtes et VIF 33](#_Toc518308949)

[Tableau 30: R, Radj, SS Model et SS residual 33](#_Toc518308950)

[Tableau 31: Paramètre Estimâtes et VIF 35](#_Toc518308951)

[Tableau 32: R,Radj, SS Model et SS residual 35](#_Toc518308952)

[Tableau 33: Number of times each predictor is referenced 37](#_Toc518308953)

[Tableau 34: Regression statistics 37](#_Toc518308954)

[Tableau 35: Model coefficients 38](#_Toc518308955)

[Tableau 36:Summary of active networks 40](#_Toc518308956)

[Tableau 37:Tests d'importance univariés pour Y\_GPER\_BDO 42](#_Toc518308957)

[Tableau 38:Paramètre Estimâtes et VIF 42](#_Toc518308958)

[Tableau 39: R,Radj, SS Model et SS residual 42](#_Toc518308959)

[Tableau 40:Number of times each predictor is referenced 44](#_Toc518308960)

[Tableau 41:Regression statistics 44](#_Toc518308961)

[Tableau 42:Model coefficients 44](#_Toc518308962)

[Tableau 43: Summary of active networks 46](#_Toc518308963)

[Tableau 44:Tests d'importance univariés pour Y\_GPER\_CDO 48](#_Toc518308964)

[Tableau 45:Paramètre Estimâtes et VIF 48](#_Toc518308965)

[Tableau 46:R,Radj, SS Model et SS residual 48](#_Toc518308966)

[Tableau 47:Number of times each predictor is referenced 50](#_Toc518308967)

[Tableau 48:Regression statistics 50](#_Toc518308968)

[Tableau 49: Model coefficients 51](#_Toc518308969)

[Tableau 50: Summary of active networks 52](#_Toc518308970)

[Tableau 51:Tests d'importance univariés pour Y\_GPER\_SS 54](#_Toc518308971)

[Tableau 52:Paramètre Estimâtes et VIF 54](#_Toc518308972)

[Tableau 53:R,Radj, SS Model et SS residual 54](#_Toc518308973)

[Tableau 54: Number of times each predictor is referenced 56](#_Toc518308974)

[Tableau 55:Regression statistics 56](#_Toc518308975)

[Tableau 56: Model coefficients 57](#_Toc518308976)

[Tableau 57: Summary of active networks 58](#_Toc518308977)

[Tableau 58: comparaison des modèles sur l'ensemble train VAR40 60](#_Toc518308978)

[Tableau 59: comparaison des modèles sur l'ensemble train VAR41 61](#_Toc518308979)

[Tableau 60: comparaison des modèles sur l'ensemble train VAR42 62](#_Toc518308980)

[Tableau 61: comparaison des modèles sur l'ensemble test VAR40 63](#_Toc518308981)

[Tableau 62: comparaison des modèles sur l'ensemble test VAR41 63](#_Toc518308982)

[Tableau 63: comparaison des modèles sur l'ensemble test VAR42 64](#_Toc518308983)

[Tableau 64: choix des modèles pour le système 64](#_Toc518308984)

[Tableau 65: T-tests 66](#_Toc518308985)

### **Résumé :**

**Dans ce devoir nous avons commencé par la création des 2 groupes : train (=1) et test (=0)**

**Dans la première partie nous avons développé des modèles sur l’ensemble train pour chaque sortie, ou nous avons utilisé (MRO, MRB, MARS, SANN) et nous avons essayé de résoudre les problèmes que nous avons rencontré (signification, multicolonéarité, points aberrantes), nous avons commencé par MRO (RÉGRESSION MULTIPLE) ensuite on a essayé de résoudre le problème de signification et multicolonéarité avec la méthode de Régression avec Backward Stepwise, et après on a essayé d’améliorer notre modèle avec l’élimination des points aberrants. Et après nous avons utilisé la méthode MARS (Multivariate Adaptative Regression Splines), Et on a terminé par la méthode de réseau de neurones.**

**Dans la deuxième partie nous avons Comparé la performance de nos modèles sur l’ensemble train et sur l’ensemble test pour chaque sortie, ou nous avons pris les trois modèles MRB, MARS, SANN (nous n'avons pas pris MRO puisque nous savons que MRB est une extension et un meilleur choix par rapport à MRO selon le devoir2), et nous avons identifié les forces et faiblesses des différentes méthodes de modélisation employées et on a terminé par une conclusion.**

**Et dans la dernière partie nous avons comparé les données d’entrée de 1990 et celles de 1991**

**Et nous avons terminé par la proposition de toute autre méthode qui on peut employer dans les questions A), B), C). Et nous avons finit par une conclusion générale.**

**Réponse**

# **A) Création des 2 groupes**

|  |  |
| --- | --- |
| **Tableau 1:les 10 premières valeurs obtenues de aléa** | **Tableau 2: ID de l'ensemble test** |

# **B) Développement des modèles sur l’ensemble train**

**Dans cette question nous avons utilisé (MRO, MRB, MARS, SANN) et nous avons essayé de résoudre les problèmes que nous avons rencontré (signification, multicolonéarité, points aberrantes)**

**Dans la première partie nous avons utilisé la RÉGRESSION MULTIPLE ensuite on a essayé de résoudre le problème de signification et multicolonéarité avec la méthode de Régression avec Backward Stepwise, et après on a essayé d’améliorer notre modèle avec l’élimination des points aberrants.**

**Et dans la deuxième partie nous avons utilisé la méthode MARS (Multivariate Adaptative Regression Splines), Et on a terminé par la méthode de réseau de neurones**

# **Var40 :Y\_GPER\_BDO**

## **RÉGRESSION MULTIPLE**

|  |  |
| --- | --- |
| **Prediction equation for:** **Y\_GPER\_BDO** = 42,7844875562-3,12866609145e-005\*FL+0,0509593713163\*"ZN"+2,39275889206\*pH+0,0411031358525\*"BDO"+0,00265271189733\*"CDO"-0,000163442493756\*SS+0,0384922216521\*VSS+0,148328762021\*SED-0,000430980307212\*COND-6,28422430411\*"pH\_P"-0,00272305034812\*"SS\_P"+0,000768732767104\*"VSS\_P"-0,0488705450994\*"SED\_P"+0,000237683665312\*"COND\_P"+3,3066404052\*"pH\_S"-0,0602488066126\*"BDO\_S"-0,0092188240828\*"CDO\_S"+0,0206650265368\*"SS\_S"-0,0556303665168\*"VSS\_S"-1,36224056348\*"SED\_S"+0,00097389024126\*"COND\_S"+0,0562365031334\*"Y\_PER\_SS\_P"-0,0665499045062\*"Y\_PER\_SED\_P"+0,609397124852\*"Y\_PER\_BDO\_S"+0,0636049193329\*"Y\_PER\_CDO\_S" | |
| Tableau 3:Tests d'importance univariés pour Y\_GPER\_BDO    Figure 1: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque qu’il y a des variables qui sont non significatifs** | |
|  | |
| Tableau 4 : Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | | **Xi** | **Y\_GPER\_BDO**  **Paramètre** | **VIF** | | **Intercept** | **42,78449** |  | | **FL** | **-0,00003** | **1,565461** | | **ZN** | **0,05096** | **1,126300** | | **pH** | **2,39276** | **5,986001** | | **BDO** | **0,04110** | **1,904251** | | **CDO** | **0,00265** | **3,338991** | | **SS** | **-0,00016** | **8,337358** | | **VSS** | **0,03849** | **7,066662** | | **SED** | **0,14833** | **6,980753** | | **COND** | **-0,00043** | **33,063255** | | **pH\_P** | **-6,28422** | **10,100794** | | **SS\_P** | **-0,00272** | **13,180982** | | **VSS\_P** | **0,00077** | **6,081508** | | **SED\_P** | **-0,04887** | **10,730493** | | **COND\_P** | **0,00024** | **31,862747** | | **pH\_S** | **3,30664** | **5,559228** | | **BDO\_S** | **-0,06025** | **2,967620** | | **CDO\_S** | **-0,00922** | **3,696718** | | **SS\_S** | **0,02067** | **4,772476** | | **VSS\_S** | **-0,05563** | **3,332351** | | **SED\_S** | **-1,36224** | **5,268729** | | **COND\_S** | **0,00097** | **11,755649** | | **Y\_PER\_SS\_P** | **0,05624** | **5,342880** | | **Y\_PER\_SED\_P** | **-0,06655** | **5,602343** | | **Y\_PER\_BDO\_S** | **0,60940** | **1,318284** | | **Y\_PER\_CDO\_S** | **0,06360** | **1,334475** |   **D’après le tableau on remarque qu’il y a le problème de multicolonéarité (il y a des Variances inflations factors** **VIF >10)**  **Pour les variables significatifs, on remarque que le signe des coefficients des variables BDO, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_BDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables pH\_P, BDO\_S, CDO\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_BDO (si Xi augmente donc Y diminué).**  **Et aussi on remarque que la variable pH\_P a eu le plus grand impact sur Y\_GPER\_BDO par unité par rapport les autres variables.** | Tableau 5: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,85 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 36,61%, il y a une partie de la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que nos modèles est moyenne bon** |
| Figure 2: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  **Et aussi que la plupart des points sont consontrer sur l’intervalle ]65 ,100[** |
| Figure 3: Analyse des résidus avec les valeurs Observes  **Nous remarquons qu'il existe des valeurs aberrantes qui influencent notre modèle prédit** |
| Figure 4: résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est moyen bon. sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue**  Figure 5:Analyse des résidus avec les valeurs Prédites | |

**D’après les remarques précédentes (VIF Variance inflation factor) on remarque qu’il y a le problème de multicolinéarité, pour résoudre se problème on peut utilisé la régression en composantes principales (ACP), la régression PLS (« Partial Least Square »), la régression RIDGE, la méthode de sélection de variables..**

**Mais comme le problème est dans les variables que ne sont pas significatif il suffit juste éliminer ces variables et pour faire ça on utilise la méthode de Régression avec Backward Stepwise.**

**On peut aussi utiliser d’autre méthode (*Forward Stepwise…*) mais dans le devoir2 on a trouvé que la méthode avec Backward Stepwise est le meilleur choix, et pour ça on a choisi cette méthode pour ce devoir.**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| Prediction equation for: **Y\_GPER\_BDO** = 39,2227110444+0,0415082508448\*"BDO"+0,00511813378411\*"CDO"-4,70667376998\*"pH\_P"+4,19948412498\*"pH\_S"-0,0577539224072\*"BDO\_S"-0,00810451589215\*"CDO\_S"-0,0350298118287\*"VSS\_S"+0,611051096854\*"Y\_PER\_BDO\_S"+0,0707164156515\*"Y\_PER\_CDO\_S" | |
| Tableau 6: Tests d'importance univariés pour Y\_GPER\_BDO    Figure 6: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque que la méthode Backward Stepwise a éliminé les variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 7: Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_BDO**  **paramétre** | **VIF** | | **Intercept** | **39,22271** |  | | **BDO** | **0,04151** | **1,7864522** | | **CDO** | **0,00512** | **2,0769254** | | **pH\_P** | **-4,70667** | **5,1850406** | | **pH\_S** | **4,19948** | **5,2205037** | | **BDO\_S** | **-0,05775** | **2,6355682** | | **CDO\_S** | **-0,00810** | **2,7449788** | | **VSS\_S** | **-0,03503** | **1,1088852** | | **Y\_PER\_BDO\_S** | **0,61105** | **1,2696199** | | **Y\_PER\_CDO\_S** | **0,07072** | **1,2350310** |   **D’après le tableau on remarque que le problème de multicolinéarité a été éliminé**  **Et aussi on remarque que le signe des coefficients des variables BDO, CDO, PH\_S Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_BDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables pH\_P, BDO\_S, CDO\_S, VSS\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_BDO (si Xi augmente donc Y diminué).**  **Et aussi que la variable pH\_P, pH\_S ont eu le plus grand impact sur Y\_GPER\_BDO par unité par rapport les autres variables.** | Tableau 8: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,85 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 38%, il y a une partie de la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèles est moyen bon** |
| Figure 7:les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  **Et aussi que la plupart des points sont consontrer sur l’intervalle ]65 ,100[** |
| Figure 8: résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** | Figure 9: Analyse des résidus avec les valeurs Observes  **Nous remarquons qu'il existe des valeurs aberrantes qui influencent notre modèle prédit** |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est moyen bon. sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue**  Figure 10: Analyse des résidus avec les valeurs Prédites | |

**la méthode Backward Stepwise a éliminé le probléme de significatifs et le probléme de multicolonéarité, et il nous reste le probléme de les points abrrentes qui ont eu une grande influence sur le modéle, et pour résoudre le probléme on a essayé d’améliorer notre modèle avec l’élimination des points aberrants.**

## **la méthode de Régression avec Backward Stepwise sans les points aberrants**

|  |  |
| --- | --- |
| Prediction equation for: Y\_GPER\_BDO = 41,2264875765-0,982727021434\*pH+0,0394861979728\*"BDO"+0,00335211459088\*"CDO"+0,000689289643663\*COND-0,0650190856681\*"BDO\_S"-0,00802929332883\*"CDO\_S"+0,0121479038596\*"SS\_S"+0,656470315663\*"Y\_PER\_BDO\_S" | |
| Tableau 9:Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_BDO**  **paramétre** | **VIF** | | **Intercept** | **41,22649** |  | | **pH** | **-0,98273** | **1,1414189** | | **BDO** | **0,03949** | **1,7842087** | | **CDO** | **0,00335** | **2,0604656** | | **COND** | **0,00069** | **1,3150428** | | **BDO\_S** | **-0,06502** | **2,6844851** | | **CDO\_S** | **-0,00803** | **2,8765589** | | **SS\_S** | **0,01215** | **1,2960765** | | **Y\_PER\_BDO\_S** | **0,65647** | **1,1269153** |   **D’après le tableau on remarque qu’il n’y a pas le problème de multicolinéarité**  **Et aussi on remarque que le signe des coefficients des variables BDO, CDO, COND, SS\_S, Y\_PER\_BDO\_S, sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_BDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables pH, BDO\_S, CDO\_S, sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_BDO (si Xi augmente donc Y diminué).**  **Et aussi que la variable pH a eu la plus grand impact sur Y\_GPER\_BDO par unité par rapport les autres variables.** | Tableau 10: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,948 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 11%, la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèles est bon** |
| Figure 11: Pareto Chart of t-Values for Coefficients  **la méthode Backward Stepwise a éliminé le problème des variables non significatifs** |
| Figure 12: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  **Et aussi que la plupart des points sont consontrer sur l’intervalle ]65 ,100[** | Figure 13: Analyse des résidus avec les valeurs Observes  **On remarque que la plupart des valeurs observes sont proches de la droite obtenue** |
| Figure 14:résidus sur échelle de probabilité gaussienne  **On remarque que les résidus respectent la loi de distribution normale (on a éliminé le problème de normalité)** | Figure 15: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,419 sur 420 ont des résidus inférieurs à 10 en valeur absolue** |

## **Modèle de Régression MARS**

0 cases with missing data were found.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MARSplines Results:

Dependent: Y\_GPER\_BDO

Independents: FL, ZN, pH, BDO, CDO, SS, VSS, SED, COND, pH\_P, SS\_P, VSS\_P, SED\_P, COND\_P, pH\_S, BDO\_S, ...

Number of terms = 15

Number of basis functions = 23

Order of interactions = 2

Penalty = 2,000000

Threshold = 0,000500

GCV error = 5,561995

Prune = Yes

**Après la régression MARS, les variables retenues dans le modèle (7 sur 25) sont FL, BDO, BDO\_S, CDO\_S, Y\_PER\_SED\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S.**

|  |  |
| --- | --- |
| Tableau 11: Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau les variables BDO(7fois), BDO\_S(4fois), Y\_PER\_BDO\_S(fois), Y\_PER\_CDO\_S(4fois) sont les plus référencées. Et les significatives.** | Tableau 12: Regression statistics    **D’après le tableau on remarque que notre modèle prédit est très bon (R2 = 89,3% et R2ajusté = 88,9%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (6,73 et 6,36)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  **Y\_GPER\_BDO** = 7,69639092370907e+001 + 5,88318336003066e-001\*max(0; Y\_PER\_BDO\_S-6,98000000000000e+001) - 1,07694073770464e+000\*max(0; 6,98000000000000e+001-Y\_PER\_BDO\_S) - 8,04139831266176e-002\*max(0; Y\_PER\_BDO\_S-6,98000000000000e+001)\*max(0; 4,18000000000000e+001-Y\_PER\_CDO\_S) + 1,68452660816721e-002\*max(0; 9,54000000000000e+001-Y\_PER\_SED\_P)\*max(0; 6,98000000000000e+001-Y\_PER\_BDO\_S) + 2,88731320436630e-002\*max(0; BDO-1,48000000000000e+002) - 9,69279751045744e-002\*max(0; 1,48000000000000e+002-BDO) + 6,58944725627016e-002\*max(0; 1,72000000000000e+002-BDO\_S) + 4,69247509719174e-004\*max(0; 1,48000000000000e+002-BDO)\*max(0; 2,12000000000000e+002-CDO\_S) - 3,36012826843113e-003\*max(0; 1,48000000000000e+002-BDO)\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) - 2,23065348577261e-002\*max(0; 1,48000000000000e+002-BDO)\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) + 3,48701485041293e-003\*max(0; 1,72000000000000e+002-BDO\_S)\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) + 2,81743639819741e-004\*max(0; 3,76620000000000e+004-FL)\*max(0; 6,98000000000000e+001-Y\_PER\_BDO\_S) + 7,30779705676608e-004\*max(0; 1,48000000000000e+002-BDO)\*max(0; BDO\_S-6,40000000000000e+001) + 5,77798918623392e-003\*max(0; 1,48000000000000e+002-BDO)\*max(0; 6,40000000000000e+001-BDO\_S) |
| Tableau 13: Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 16: : les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 17: : Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon. sur les variables prédites,418 sur 421 ont des résidus inférieurs à 10 en valeur absolue** |

**Les graphiques qui identifient les noeuds employés dans le modèle :**

|  |  |
| --- | --- |
| Figure 18: Y\_GPER\_BDO avec FL | Figure 19: Y\_GPER\_BDO avec BDO |
| Figure 20: Y\_GPER\_BDO avec BDO\_S | Figure 21: Y\_GPER\_BDO avec CDO\_S |
| Figure 22:Y\_GPER\_BDO avec Y\_PER\_SED\_P | Figure 23: Y\_GPER\_BDO avec Y\_PER\_BDO\_S |

****

Figure 24: Y\_GPER\_BDO avec Y\_PER\_CDO\_S

## **Réseaux de neurones (SANN**

**Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs**

|  |
| --- |
| Tableau 14: Summary of active networks |

|  |  |
| --- | --- |
| Figure 25:les valeurs Prédites modèle 1 avec les valeurs Observes  **R²= 0,71717458**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 26: les valeurs Prédites modèle 2 avec les valeurs Observes  **R²= 0,79649223**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

**D’après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le deuxièmes réseau (R²=0,7964) meilleur par rapport au premier (R²=0,7117)**

|  |  |
| --- | --- |
| Figure 27: les valeurs Prédites avec les valeurs Observes | **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées) pour les deux modèles** |
| Figure 28: Analyse des résidus avec les valeurs Prédites modèle 1  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,418 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | Figure 29: Analyse des résidus avec les valeurs Prédites modèle 2  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,418 sur 421 ont des résidus inférieurs à 8 en valeur absolue** |

**D’après les graphes nous avons confirmé notre remarque président (le second réseau est meilleure par rapport au premier)**

# **VAR41 : Y\_GPERF\_CDO**

## **RÉGRESSION MULTIPLE**

|  |  |
| --- | --- |
| **Prediction equation for:** **Y\_GPER\_CDO** = 21,8360819469-4,47145181572e-005\*FL+0,0789680434862\*"ZN"+0,597881276564\*pH+0,00331871223286\*"BDO"+0,0414525918944\*"CDO"-0,000246977246536\*SS+0,0403768025774\*VSS-0,0981110995718\*SED+0,00214811787549\*COND+1,09377167641\*"pH\_P"-0,00143686026081\*"SS\_P"+0,0180544603209\*"VSS\_P"+0,13244490338\*"SED\_P"-0,00123175741227\*"COND\_P"-1,43801742424\*"pH\_S"-0,0299431421307\*"BDO\_S"-0,0548663215201\*"CDO\_S"-0,00633944450595\*"SS\_S"-0,0497418376899\*"VSS\_S"+0,699365300633\*"SED\_S"-0,00174093995295\*"COND\_S"+0,0754661150403\*"Y\_PER\_SS\_P"-0,0106782059074\*"Y\_PER\_SED\_P"+0,175196451299\*"Y\_PER\_BDO\_S"+0,597030768279\*"Y\_PER\_CDO\_S" | |
| Tableau 15: Tests d'importance univariés pour Y\_GPER\_CDO    Figure 30: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque qu’il y a des variables qui sont non significatifs** | |
|  | |
| Tableau 16 : Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_CDO**  **paramétre** | **VIF** | | **Intercept** | **21,83608** |  | | **FL** | **-0,00004** | **1,565461** | | **ZN** | **0,07897** | **1,126300** | | **pH** | **0,59788** | **5,986001** | | **BDO** | **0,00332** | **1,904251** | | **CDO** | **0,04145** | **3,338991** | | **SS** | **-0,00025** | **8,337358** | | **VSS** | **0,04038** | **7,066662** | | **SED** | **-0,09811** | **6,980753** | | **COND** | **0,00215** | **33,063255** | | **pH\_P** | **1,09377** | **10,100794** | | **SS\_P** | **-0,00144** | **13,180982** | | **VSS\_P** | **0,01805** | **6,081508** | | **SED\_P** | **0,13244** | **10,730493** | | **COND\_P** | **-0,00123** | **31,862747** | | **pH\_S** | **-1,43802** | **5,559228** | | **BDO\_S** | **-0,02994** | **2,967620** | | **CDO\_S** | **-0,05487** | **3,696718** | | **SS\_S** | **-0,00634** | **4,772476** | | **VSS\_S** | **-0,04974** | **3,332351** | | **SED\_S** | **0,69937** | **5,268729** | | **COND\_S** | **-0,00174** | **11,755649** | | **Y\_PER\_SS\_P** | **0,07547** | **5,342880** | | **Y\_PER\_SED\_P** | **-0,01068** | **5,602343** | | **Y\_PER\_BDO\_S** | **0,17520** | **1,318284** | | **Y\_PER\_CDO\_S** | **0,59703** | **1,334475** |   **D’après le tableau on remarque qu’il y a le problème de multicolonéarité (il y a des VIF >10)**  **Pour les variables significatifs, on remarque que le signe des coefficients des variables CDO, Y\_PER\_SS\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_CDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables BDO\_S, CDO\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_CDO (si Xi augmente donc Y diminué).**  **Et aussi on remarque que la variable Y\_PER\_CDO\_S a eu le plus grand impact sur Y\_GPER\_CDO par unité par rapport les autres variables.** | Tableau 17: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,92 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 17%, la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèle est bon** |
| Figure 31: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |
| Figure 32: Analyse des résidus avec les valeurs Observes  **Nous remarquons qu'il existe des valeurs aberrantes qui influencent notre modèle prédit** |
| Figure 33: résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est moyen bon. sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue**  Figure 34:Analyse des résidus avec les valeurs Prédites | |

**D’après VIF on remarque qu’il y a le problème de multicolinéarité pour le résoudre se problème on peut utilisé la régression en composantes principales (ACP), la régression PLS (« Partial Least Square »), la régression RIDGE, la méthode de sélection de variables..**

**Mais comme le problème est dans les variables que ne sont pas significatif il suffit juste éliminer ces variables et pour faire ça on utilise la méthode de Régression avec Backward Stepwise.**

**On peut aussi utiliser d’autre méthode (*Forward Stepwise…*) mais dans le devoir2 on a trouvé que la méthode avec Backward Stepwise est le meilleur choix, et pour ça on a l’utilisé dans ce devoir.**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| **Prediction equation for: Y\_GPER\_CDO** = 19,3652006125+0,0427269579162\*"CDO"+0,0308786584924\*VSS-0,0264583045016\*"BDO\_S"-0,0563813164067\*"CDO\_S"+0,0547674665868\*"Y\_PER\_SS\_P"+0,16870545371\*"Y\_PER\_BDO\_S"+0,599362400742\*"Y\_PER\_CDO\_S" | |
| Tableau 18: Tests d'importance univariés pour Y\_GPER\_CDO    Figure 35: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque que la méthode Backward Stepwise a éliminé le problème des variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 19: Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_CDO**  **paramétre** | **Variance**  **VIF** | | **Intercept** | **19,36520** |  | | **CDO** | **0,04273** | **2,0424063** | | **VSS** | **0,03088** | **1,4300429** | | **BDO\_S** | **-0,02646** | **2,2254816** | | **CDO\_S** | **-0,05638** | **2,8707497** | | **Y\_PER\_SS\_P** | **0,05477** | **1,3089060** | | **Y\_PER\_BDO\_S** | **0,16871** | **1,2457214** | | **Y\_PER\_CDO\_S** | **0,59936** | **1,2097871** |   **D’après le tableau on remarque que le problème de multicolinéarité a été éliminé**  **Et aussi on remarque que le signe des coefficients des variables CDO,VSS, Y\_PER\_SS\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_CDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables BDO\_S, CDO\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_CDO (si Xi augmente donc Y diminué).**  **Et aussi que la variable Y\_PER\_CDO\_S a eu le plus grand impact sur Y\_GPER\_CDO par unité par rapport les autres variables.** | Tableau 20: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,923 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 17,3%, la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèle est bon** |
| Figure 36: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  **Et aussi que la plupart des points sont consontrer sur l’intervalle ]65 ,100[** |
| Figure 37:résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** | Figure 38: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est moyen bon. sur les variables prédites,419 sur 421 ont des résidus inférieurs à 10 en valeur absolue** |

**Modèle de Régression MARS**

0 cases with missing data were found.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MARSplines Results:

Dependent: Y\_GPER\_BDO

Independents: FL, ZN, pH, BDO, CDO, SS, VSS, SED, COND, pH\_P, SS\_P, VSS\_P, SED\_P, COND\_P, pH\_S, BDO\_S, ...

Number of terms = 15

Number of basis functions = 23

Order of interactions = 2

Penalty = 2,000000

Threshold = 0,000500

GCV error = 5,561995

Prune = Yes

**Après la régression MARS, les variables retenues dans le modèle (7 sur 25) sont CDO, BDO\_S, CDO\_S,Y\_PER\_SS, Y\_PER\_SED\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S.**

|  |  |
| --- | --- |
| Tableau 21: Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau les variables CDO (6fois), Y\_PER\_CDO\_S(8fois) sont les plus référencées. Et les significatives.** | Tableau 22: Regression statistics    **D’après le tableau on remarque que notre modèle prédit est très bon (R2 = 94,4% et R2ajusté = 94,2%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (8,36 et 8,12)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  **Y\_GPER\_CDO** = 9,28929481798562e+001 + 1,07374786166846e+000\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) - 3,81896587704122e-001\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) - 1,72274947612316e-002\*max(0; Y\_PER\_SS\_P-7,11000000000000e+001)\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) - 1,81070856767918e-002\*max(0; 7,11000000000000e+001-Y\_PER\_SS\_P)\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) + 4,17048108721013e-002\*max(0; CDO-4,78000000000000e+002) - 6,91112860013462e-002\*max(0; 4,78000000000000e+002-CDO) - 8,56476260857855e-002\*max(0; CDO\_S-8,00000000000000e+001) - 5,60162556236426e-002\*max(0; 6,98000000000000e+001-Y\_PER\_BDO\_S)\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) + 4,59836816047356e-004\*max(0; 9,90000000000000e+001-BDO\_S)\*max(0; CDO\_S-8,00000000000000e+001) + 1,56825292430057e-004\*max(0; 4,78000000000000e+002-CDO)\*max(0; Y\_PER\_SS\_P-3,40000000000000e+001) - 1,10069967017266e-003\*max(0; 4,78000000000000e+002-CDO)\*max(0; 3,40000000000000e+001-Y\_PER\_SS\_P) + 6,35246612880811e-004\*max(0; 4,78000000000000e+002-CDO)\*max(0; 8,48000000000000e+001-Y\_PER\_SED\_P) + 2,39652236071706e-003\*max(0; CDO\_S-2,96000000000000e+002)\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) - 1,49023617291979e-003\*max(0; 2,96000000000000e+002-CDO\_S)\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) - 1,18849112925676e-003\*max(0; CDO-8,10000000000000e+001)\*max(0; Y\_PER\_CDO\_S-5,83000000000000e+001) |
| Tableau 23: Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 39: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 40: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon. sur les variables prédites,420 sur 421 ont des résidus inférieurs à 10 en valeur absolue** |

**Les graphiques qui identifient les noeuds employés dans le modèle :**

|  |  |
| --- | --- |
| Figure 41: Y\_GPER\_CDO avec CDO | Figure 42: Y\_GPER\_CDO avec BDO\_S |
| Figure 43: Y\_GPER\_CDO avec CDO\_S | Figure 44: Y\_GPER\_CDO avec Y\_PER\_SS\_P |
| Figure 45: Y\_GPER\_CDO avec Y\_PER\_SED\_P | Figure 46: Y\_GPER\_CDO avec Y\_PER\_BDO\_S |

****

Figure 47: Y\_GPER\_CDO avec Y\_PER\_CDO\_S

## **Réseaux de neurones (SANN)**

**Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs**

|  |
| --- |
| Tableau 24: Summary of active networks |

|  |  |
| --- | --- |
| Figure 48: les valeurs Prédites modèle 1 avec les valeurs Observes  **R²= 0,94028637**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 49: les valeurs Prédites modèle 2 avec les valeurs Observes  **R²= 0,94073854**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

|  |  |
| --- | --- |
| Figure 50: les valeurs Prédites avec les valeurs Observes | **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées) pour les deux modèles** |
| Figure 51: Analyse des résidus avec les valeurs Prédites modèle 1  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,419 sur 421 ont des résidus inférieurs à 8 en valeur absolue** | Figure 52: Analyse des résidus avec les valeurs Prédites modèle 2  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,419 sur 421 ont des résidus inférieurs à 8 en valeur absolue** |

**D’après les graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le deuxièmes réseau (R²=0,9407) meilleur par rapport au premier (R²=0,9402)**

# **VAR42 : Y\_GPERF\_SS**

## **RÉGRESSION MULTIPLE**

|  |  |
| --- | --- |
| **Prediction equation for:** **Y\_GPER\_SS** = 27,4147180021-2,01301116685e-005\*FL+0,265814522202\*"ZN"+1,02965361759\*pH-0,00190594816331\*"BDO"+0,014734331722\*"CDO"+0,000595833855206\*SS-0,020751334989\*VSS+0,36664611104\*SED+0,00143074759498\*COND-6,05142712649\*"pH\_P"+0,000909921498375\*"SS\_P"+0,0211792250563\*"VSS\_P"-0,261246687754\*"SED\_P"-0,000524747855741\*"COND\_P"+7,75864960496\*"pH\_S"-0,0580518317973\*"BDO\_S"-0,0196244245622\*"CDO\_S"+0,0620477542533\*"SS\_S"+0,00477753009798\*"VSS\_S"-1,44478764923\*"SED\_S"-0,00172245701304\*"COND\_S"+0,19744779431\*"Y\_PER\_SS\_P"-0,129299037642\*"Y\_PER\_SED\_P"+0,31547608992\*"Y\_PER\_BDO\_S"+0,233847860574\*"Y\_PER\_CDO\_S" | |
| Tableau 25: Tests d'importance univariés pour Y\_GPERF\_SS    Figure 53: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque qu’il y a des variables qui sont non significatifs** | |
|  | |
| Tableau 26: Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPERF\_SS**  **paramétre** | **VIF** | | **Intercept** | **27,41472** |  | | **FL** | **-0,00002** | **1,565461** | | **ZN** | **0,26581** | **1,126300** | | **pH** | **1,02965** | **5,986001** | | **BDO** | **-0,00191** | **1,904251** | | **CDO** | **0,01473** | **3,338991** | | **SS** | **0,00060** | **8,337358** | | **VSS** | **-0,02075** | **7,066662** | | **SED** | **0,36665** | **6,980753** | | **COND** | **0,00143** | **33,063255** | | **pH\_P** | **-6,05143** | **10,100794** | | **SS\_P** | **0,00091** | **13,180982** | | **VSS\_P** | **0,02118** | **6,081508** | | **SED\_P** | **-0,26125** | **10,730493** | | **COND\_P** | **-0,00052** | **31,862747** | | **pH\_S** | **7,75865** | **5,559228** | | **BDO\_S** | **-0,05805** | **2,967620** | | **CDO\_S** | **-0,01962** | **3,696718** | | **SS\_S** | **0,06205** | **4,772476** | | **VSS\_S** | **0,00478** | **3,332351** | | **SED\_S** | **-1,44479** | **5,268729** | | **COND\_S** | **-0,00172** | **11,755649** | | **Y\_PER\_SS\_P** | **0,19745** | **5,342880** | | **Y\_PER\_SED\_P** | **-0,12930** | **5,602343** | | **Y\_PER\_BDO\_S** | **0,31548** | **1,318284** | | **Y\_PER\_CDO\_S** | **0,23385** | **1,334475** |   **D’après le tableau on remarque qu’il y a le problème de multicolonéarité (il y a des Variances inflations factors** **VIF >10)**  **Pour les variables significatifs, on remarque que le signe des coefficients des variables ZN, CDO, SS\_S, Y\_PER\_SS\_P, Y\_PER\_BDO\_S,Y\_PER\_COD\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPERF\_SS (si Xi augmente donc Y augmente). Et le signe des coefficients des variables pH\_P, BDO\_S, CDO\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPEF\_SS (si Xi augmente donc Y diminué).**  **Et aussi on remarque que la variable**  **Y\_PER\_BDO\_S a eu le plus grand impact sur Y\_GPER\_BDO par unité par rapport les autres variables.** | Tableau 27: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination < 𝟎, 𝟓 (𝑹=0,63< 𝟎, 𝟕), on remarque que on a une mauvaise corrélation entre Xi et Y ,notre modèle est mauvais.** |
| Figure 54: les valeurs Prédites avec les valeurs Observes  **On remarque que notre modèle n’est pas bon, et aussi que la plupart des points sont consontrer sur l’intervalle ]75 ,100[** |
| Figure 55: Analyse des résidus avec les valeurs Observes  **Nous remarquons qu'il existe des valeurs aberrantes qui influencent notre modèle prédit** |
| Figure 56: résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu n’est pas bon. sur les variables prédites,417 sur 421 ont des résidus inférieurs à 20 en valeur absolue**  Figure 57: Analyse des résidus avec les valeurs Prédites | |

**D’après VIF on remarque qu’il y a le problème de multicolinéarité, pour résoudre se problème on peut utilisé la régression en composantes principales (ACP), la régression PLS (« Partial Least Square »), la régression RIDGE, la méthode de sélection de variables..**

**Mais comme le problème est dans les variables que ne sont pas significatif il suffit juste éliminer ces variables et pour faire ça on utilise la méthode de Régression avec Backward Stepwise.**

**On peut aussi utiliser d’autre méthode (*Forward Stepwise…*) mais dans le devoir2 on a trouvé que la méthode avec Backward Stepwise est le meilleur choix, et pour ça on a l’utilisé dans ce devoir.**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| **Prediction equation for:** **Y\_GPER\_SS** = 44,5728981017+0,296859775556\*"ZN"+0,0177503398727\*"CDO"-0,0603784480986\*"BDO\_S"-0,02608288429\*"CDO\_S"+0,0527857916844\*"SS\_S"+0,191107908001\*"Y\_PER\_SS\_P"-0,104968416465\*"Y\_PER\_SED\_P"+0,32705487803\*"Y\_PER\_BDO\_S"+0,249939645544\*"Y\_PER\_CDO\_S" | |
| Tableau 28: Tests d'importance univariés pour Y\_GPEF\_SS    **D’après le tableau (p-valeu) on remarque que la méthode Backward Stepwise a éliminé le problème des variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 29: Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_SS**  **paramétre** | **VIF** | | **Intercept** | **44,57290** |  | | **ZN** | **0,29686** | **1,0196591** | | **CDO** | **0,01775** | **2,0306544** | | **BDO\_S** | **-0,06038** | **2,3285832** | | **CDO\_S** | **-0,02608** | **2,8276104** | | **SS\_S** | **0,05279** | **1,5325138** | | **Y\_PER\_SS\_P** | **0,19111** | **1,6657532** | | **Y\_PER\_SED\_P** | **-0,10497** | **1,6775825** | | **Y\_PER\_BDO\_S** | **0,32705** | **1,2628387** | | **Y\_PER\_CDO\_S** | **0,24994** | **1,2258142** |   **D’après le tableau on remarque que le problème de multicolinéarité a été éliminé**  **Et aussi on remarque que le signe des coefficients des variables ZN, CDO, SS\_S, Y\_PER\_SS\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPERF\_SS (si Xi augmente donc Y augmente). Et le signe des coefficients des variables BDO\_S, CDO\_S, Y\_PER\_SED\_P sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPERF\_SS (si Xi augmente donc Y diminué).** | Tableau 30: R, Radj, SS Model et SS residual    **D’après le coefficient de détermination < 𝟎, 𝟓 (𝑹=0,62< 𝟎, 𝟕), on remarque que on a une mauvaise corrélation entre Xi et Y notre modèle est mauvais.** |
| Figure 58 :les valeurs Prédites avec les valeurs Observes  **On remarque que notre modèle n’est pas bon**  **On remarque que la plupart des valeurs prédites ne sont pas proches de la droite obtenue** |
| Figure 59: résidus sur échelle de probabilité gaussienne  **Nous remarquons qu'il existe des valeurs aberrantes qui affectent la normalité de notre modèle** | Figure 60: Analyse des résidus avec les valeurs Observes  **On remarque que notre modèle n’est pas bon** |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu n’est pas bon (il y a une grande marge d’erreur). sur les variables prédites,417 sur 421 ont des résidus inférieurs à 20 en valeur absolue**  Figure 61: Analyse des résidus avec les valeurs Prédites | |

**la méthode Backward Stepwise a éliminé le probléme de significatifs et le probléme de multicolonéarité, et il nous reste le probléme de les points abrrentes qui ont eu une grande influence sur le modéle, et pour résoudre le probléme on a essayé d’améliorer notre modèle avec l’élimination des points aberrants.**

## **la méthode de Régression avec Backward Stepwise sans les points aberrants**

|  |  |
| --- | --- |
| **Prediction equation for**: **Y\_GPER\_SS** = 38,9387240039+0,179051991641\*"ZN"+0,0127655639561\*"CDO"-0,0621837674627\*VSS+2,05219959726\*"pH\_S"-0,060141933307\*"BDO\_S"-0,0168018373179\*"CDO\_S"+0,0373585948913\*"SS\_S"+0,153643447808\*"Y\_PER\_SS\_P"-0,0890770881143\*"Y\_PER\_SED\_P"+0,372376015121\*"Y\_PER\_BDO\_S"+0,133056523489\*"Y\_PER\_CDO\_S" | |
| Tableau 31: Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPERF\_SS**  **Paramétre** | **VIF** | | **Intercept** | **38,93872** |  | | **ZN** | **0,17905** | **1,0393765** | | **CDO** | **0,01277** | **2,1644559** | | **VSS** | **-0,06218** | **2,0187304** | | **pH\_S** | **2,05220** | **1,1315970** | | **BDO\_S** | **-0,06014** | **2,3949554** | | **CDO\_S** | **-0,01680** | **2,9482235** | | **SS\_S** | **0,03736** | **1,9455894** | | **Y\_PER\_SS\_P** | **0,15364** | **2,2180817** | | **Y\_PER\_SED\_P** | **-0,08908** | **1,7346957** | | **Y\_PER\_BDO\_S** | **0,37238** | **1,2966025** | | **Y\_PER\_CDO\_S** | **0,13306** | **1,2771850** |   **D’après le tableau on remarque qu’il n’y a pas le problème de multicolinéarité**  **Et aussi on remarque que le signe des coefficients des variables ZN, CDOpH\_s, SS\_S, Y\_PER\_SS\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S, sont positive cela implique que ces variables évalué dans le même sens que Y\_GPERF\_SS (si Xi augmente donc Y augmente). Et le signe des coefficients des variables VSS, BDO\_S, CDO\_S, Y\_PER\_SED\_P sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPERF\_SS (si Xi augmente donc Y diminué).**  **Et aussi que la variable pH\_S a eu la plus grand impact sur Y\_GPERF\_SS par unité par rapport les autres variables.** | Tableau 32: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,799 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. Ce que qui montre que notre modèles est moyen bon.** |
| Figure 62: Analyse des résidus avec les valeurs Observes  **On remarque que la plupart des valeurs observes sont proches de la droite obtenue** |
| Figure 63: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  **on remarque que la plupart des points sont consontrer sur l’intervalle ]70 ,100[** | Figure 64: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement moyen bon. sur les variables prédites,416 sur 417 ont des résidus inférieurs à 15 en valeur absolue** |
| Figure 65: résidus sur échelle de probabilité gaussienne  **On remarque que les résidus respectent la loi de distribution normale (on a éliminé le problème de normalité)** |  |

**Modèle de Régression MARS**

0 cases with missing data were found.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MARSplines Results:

Dependent: Y\_GPER\_SS

Independents: FL, ZN, pH, BDO, CDO, SS, VSS, SED, COND, pH\_P, SS\_P, VSS\_P, SED\_P, COND\_P, pH\_S, BDO\_S, ...

Number of terms = 14

Number of basis functions = 20

Order of interactions = 2

Penalty = 2,000000

Threshold = 0,000500

GCV error = 33,217544

Prune = Yes

**Après la régression MARS, les variables retenues dans le modèle (6 sur 25) sont SS, COND\_P,BDO\_S, Y\_PER\_SS\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S.**

|  |  |
| --- | --- |
| Tableau 33: Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau les variables SS, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont les plus référencées. Et les significatives.** | Tableau 34: Regression statistics    **D’après le tableau on remarque que notre modèle prédit est moyen bon (R2 = 59,6% et R2ajusté = 58,29%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (8,5 et 6,57)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  **Y\_GPER\_SS** = 8,17005709243861e+001 + 3,24185559744647e-001\*max(0; Y\_PER\_CDO\_S-4,18000000000000e+001) + 7,47142135783460e-001\*max(0; 4,18000000000000e+001-Y\_PER\_CDO\_S) - 1,18276294697756e-001\*max(0; 7,19000000000000e+001-Y\_PER\_SS\_P)\*max(0; 4,18000000000000e+001-Y\_PER\_CDO\_S) + 8,83163313541901e-003\*max(0; SS-2,44000000000000e+002) - 1,15557399677833e-001\*max(0; 2,44000000000000e+002-SS) + 5,81543371443944e-003\*max(0; 2,44000000000000e+002-SS)\*max(0; Y\_PER\_BDO\_S-7,50000000000000e+001) + 2,69573391129581e-003\*max(0; 2,44000000000000e+002-SS)\*max(0; 7,60000000000000e+001-BDO\_S) - 5,56816445589623e-003\*max(0; 2,44000000000000e+002-SS)\*max(0; 5,83000000000000e+001-Y\_PER\_CDO\_S) - 1,77140714645494e-002\*max(0; Y\_PER\_BDO\_S-6,98000000000000e+001)\*max(0; Y\_PER\_CDO\_S-4,18000000000000e+001) - 1,57762822979340e-002\*max(0; 6,98000000000000e+001-Y\_PER\_BDO\_S)\*max(0; Y\_PER\_CDO\_S-4,18000000000000e+001) + 9,11228074663858e-001\*max(0; Y\_PER\_BDO\_S-7,95000000000000e+001) - 6,32710448449902e-003\*max(0; COND\_P-1,64500000000000e+003)\*max(0; 7,95000000000000e+001-Y\_PER\_BDO\_S) + 6,22691701736990e-002\*max(0; 1,83000000000000e+002-BDO\_S) |
| Tableau 35: Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 66: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 67: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon. sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue** |

**Les graphiques qui identifient les noeuds employés dans le modèle :**

|  |  |
| --- | --- |
| Figure 68: Y\_GPER\_SS avec SS | Figure 69: Y\_GPER\_SS avec COND\_P |
| Figure 70: Y\_GPER\_SS avec BDO\_S | Figure 71: Y\_GPER\_SS avec Y\_PER\_SS\_P |
| Figure 72: Y\_GPER\_SS avec Y\_PER\_BDD\_S | Figure 73: Y\_GPER\_SS avec Y\_PER\_CDO\_S |

**Réseaux de neurones (SANN)**

**Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs**

|  |
| --- |
| Tableau 36:Summary of active networks |

|  |  |
| --- | --- |
| Figure 74: es valeurs Prédites modèle 1 avec les valeurs Observes  **R²= 0,62044519**  **On remarque qu’il y a des valeurs prédites ne sont pas proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 75:les valeurs Prédites modèle 2 avec les valeurs Observes  **R²= 0,62928732**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

**D’après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le deuxièmes réseau (R²=0,629) meilleur par rapport au premier (R²=0,620)**

|  |  |
| --- | --- |
| Figure 76: les valeurs Prédites avec les valeurs Observes | **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées) pour les deux modèles** |
| Figure 77 :Analyse des résidus avec les valeurs Prédites modèle 1  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu n’est pas bon. sur les variables prédites,418 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | Figure 78:Analyse des résidus avec les valeurs Prédites modèle 2  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,421 sur 421 ont des résidus inférieurs à 10 en valeur absolue** |

**D’après les graphes nous avons confirmé notre remarque président (le second réseau est meilleure par rapport au premier)**

# **C) Comparer la performance de nos modèles sur l’ensemble train et sur l’ensemble test.**

**Au début nous avons employée nos modèles sur l’ensemble test (pour chaque Y ), et ensuite nous avons vérifié nos modéles sur l’ensemble train et l’ensemble test.**

# **Var40 :Y\_GPER\_BDO sur l’ensemble test**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| Prediction equation for: Y\_GPER\_BDO = 31,075525566+0,0454454915062\*"BDO"+0,00279547922547\*"COND\_P"-0,0906284933275\*"BDO\_S"-0,00289142245115\*"COND\_S"+0,688996870066\*"Y\_PER\_BDO\_S"+0,0482335839965\*"Y\_PER\_CDO\_S" | |
| Tableau 37:Tests d'importance univariés pour Y\_GPER\_BDO    Figure 79: Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque que la méthode Backward Stepwise a éliminé le problème des variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 38:Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | Y\_GPER\_BDO  paramétre | Variance  **VIF** | | Intercept | 31,07553 |  | | BDO | 0,04545 | 1,486881 | | COND\_P | 0,00280 | 10,655426 | | BDO\_S | -0,09063 | 1,599238 | | COND\_S | -0,00289 | 10,856529 | | Y\_PER\_BDO\_S | 0,68900 | 1,751876 | | Y\_PER\_CDO\_S | 0,04823 | 1,642240 |   **D’après le tableau on remarque qu’il y a le problème de multicolinéarité**  **Et aussi on remarque que le signe des coefficients des variables BDO, COND\_P, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_BDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables BDO\_S, COND\_S sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_BDO (si Xi augmente donc Y diminué).** | Tableau 39: R,Radj, SS Model et SS residual    **D’après le coefficient de détermination**  > 𝟎, 𝟓 (𝑹=0,96 > 𝟎, 𝟕)**, on remarque que on a une très bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 7%, il y a une petite partie de la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèles est très bon** |
| Figure 80: les valeurs Prédites avec les valeurs Observes  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)**  Et aussi **que la plupart des points sont consontrer sur l’intervalle ]70 ,100]** |
| Figure 81:résidus sur échelle de probabilité gaussienne  **On remarque que notre modèle suit la loi normale** | Figure 82: Analyse des résidus avec les valeurs Observes |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 6 en valeur absolue**  Figure 83: Analyse des résidus avec les valeurs Prédites | |

## **Modèle de Régression MARS**

**Après la régression MARS, les variables retenues dans le modèle (5 sur 25) sont BDO, BDO\_S, COND\_S, Y\_PER\_SED\_P, Y\_PER\_BDO\_S.**

|  |  |
| --- | --- |
| Tableau 40:Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau les variables BDO(5fois), BDO\_S(4fois), Y\_PER\_BDO\_S(6fois) sont les plus référencées. Et les significatives.** | Tableau 41:Regression statistics    **D’après le tableau on remarque que notre modèle prédit est très bon (R2 = 99,4% et R2ajusté = 99,3%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (6,97 et 6,95)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  **Y\_GPER\_BDO** = 9,15013844235010e+001 + 6,38525058047884e-001\*max(0; Y\_PER\_BDO\_S-7,87000000000000e+001) - 4,73682349746185e-001\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) + 2,03570349031628e-002\*max(0; BDO-1,52000000000000e+002) - 6,12862319928229e-002\*max(0; 1,52000000000000e+002-BDO) - 5,37847326850605e-002\*max(0; BDO\_S-3,40000000000000e+001) - 3,17965867840417e-002\*max(0; BDO\_S-1,22000000000000e+002)\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) + 1,00028852519149e-002\*max(0; 1,22000000000000e+002-BDO\_S)\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) + 1,77782730546644e-003\*max(0; BDO-2,02000000000000e+002)\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) - 7,14788057195776e-003\*max(0; 2,02000000000000e+002-BDO)\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) - 2,29262519381417e-004\*max(0; 2,53000000000000e+002-BDO)\*max(0; BDO\_S-3,40000000000000e+001) + 2,13294781467418e-003\*max(0; 1,19000000000000e+003-COND\_S) + 8,03028734120731e-003\*max(0; 1,19000000000000e+003-COND\_S)\*max(0; Y\_PER\_SED\_P-9,50000000000000e+001) |
| Tableau 42:Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 84:les valeurs Prédites avec les valeurs Observes  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 85:Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est trés bon. sur les variables prédites,106 sur 106 ont des résidus inférieurs à 2,5 en valeur absolue** |

## **Réseaux de neurones (SANN)**

**Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs**

|  |
| --- |
| Tableau 43: Summary of active networks |

|  |  |
| --- | --- |
| Figure 86:les valeurs Prédites modèle 1 avec les valeurs Observes  **R²= 0,98787540**  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 87: les valeurs Prédites modèle 2 avec les valeurs Observes  **R²= 0,98488598**  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

**D’après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le premier réseau (R²=0,987) meilleur par rapport au deuxièmes (R²=0,984)**

|  |  |
| --- | --- |
|  |  |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 1,5 en valeur absolue** | **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 3 en valeur absolue** |

# **VAR41 : Y\_GPERF\_CDO sur l’ensemble test**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| **Prediction equation for:** **Y\_GPER\_CDO** = 13,7512021611-0,000105067344117\*FL+0,0364471500794\*"CDO"-0,44985067033\*"SED\_P"-0,0647340354805\*"CDO\_S"+3,48123153308\*"SED\_S"+0,192480617735\*"Y\_PER\_SED\_P"+0,160343433754\*"Y\_PER\_BDO\_S"+0,602921128661\*"Y\_PER\_CDO\_S" | |
| Tableau 44:Tests d'importance univariés pour Y\_GPER\_CDO    Figure 88:Pareto Chart of t-Values for Coefficients  **D’après le tableau et le graphe (p-valeu) on remarque que la méthode Backward Stepwise a éliminé le problème des variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 45:Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | Y\_GPER\_CDO  paramétre | **VIF** | | Intercept | 13,75120 |  | | FL | -0,00011 | 1,2285428 | | CDO | 0,03645 | 1,9927868 | | SED\_P | -0,44985 | 2,1310811 | | CDO\_S | -0,06473 | 1,7593550 | | SED\_S | 3,48123 | 5,1553036 | | Y\_PER\_SED\_P | 0,19248 | 5,1767003 | | Y\_PER\_BDO\_S | 0,16034 | 1,7091242 | | Y\_PER\_CDO\_S | 0,60292 | 1,7476916 |   **D’après le tableau on remarque que le problème de multicolinéarité a été éliminé**  **Et aussi on remarque que le signe des coefficients des variables CDO, SED\_S, Y\_PER\_SED\_S, Y\_PER\_BDO\_S, Y\_PER\_CDO\_S sont positive cela implique que ces variables évalué dans le même sens que Y\_GPER\_CDO (si Xi augmente donc Y augmente). Et le signe des coefficients des variables SED\_P, CDO\_S, sont négative, implique que les variables évaluées dans le sens inverse de Y\_GPER\_CDO (si Xi augmente donc Y diminué).** | Tableau 46:R,Radj, SS Model et SS residual    **D’après le coefficient de détermination**  > 𝟎, 𝟓 (𝑹=0,96 > 𝟎, 𝟕)**, on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 6,8%, la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèle est très bon** |
| Figure 89:les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |
| Figure 90:résidus sur échelle de probabilité gaussienne  **On remarque que notre modèle suit la loi normale** | Figure 91: Analyse des résidus avec les valeurs Observes |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 8 en valeur absolue**  Figure 92:Analyse des résidus avec les valeurs Prédites | |

**Modèle de Régression MARS**

0 cases with missing data were found.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MARSplines Results:

Dependent: Y\_GPER\_CDO

Independents: FL, ZN, pH, BDO, CDO, SS, VSS, SED, COND, pH\_P, SS\_P, VSS\_P, SED\_P, COND\_P, pH\_S, BDO\_S, ...

Number of terms = 17

Number of basis functions = 25

Order of interactions = 2

Penalty = 2,000000

Threshold = 0,000500

GCV error = 1,006549

Prune = Yes

**Après la régression MARS, les variables retenues dans le modèle (7 sur 25) sont CDO, VSS, SED, SED\_P,CDO\_S,VSS\_S, Y\_PER\_CDO\_S.**

|  |  |
| --- | --- |
| Tableau 47:Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau les variables CDO(7fois), CDO\_S(7fois), Y\_PER\_CDO\_S(7fois) sont les plus référencées. Et les significatives.** | Tableau 48:Regression statistics    **D’après le tableau on remarque que notre modèle prédit est très bon (R2 = 99,4% et R2ajusté = 99,3%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (9,74 et 9,72)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  Y\_GPER\_CDO = 6,30073036756078e+001 + 3,94884765557069e-001\*max(0; Y\_PER\_CDO\_S-5,29000000000000e+001) - 5,45890443940775e-001\*max(0; 5,29000000000000e+001-Y\_PER\_CDO\_S) + 7,61260601014280e-002\*max(0; CDO-3,38000000000000e+002) - 1,31637836925887e-001\*max(0; 3,38000000000000e+002-CDO) - 1,12019070475210e-001\*max(0; CDO\_S-2,72000000000000e+002) + 1,55203279878421e-001\*max(0; 2,72000000000000e+002-CDO\_S) - 3,02458151153637e-002\*max(0; 7,18000000000000e+001-VSS\_S)\*max(0; 5,29000000000000e+001-Y\_PER\_CDO\_S) - 2,32727939590156e-003\*max(0; CDO-5,38000000000000e+002)\*max(0; Y\_PER\_CDO\_S-5,29000000000000e+001) + 1,84488233240850e-003\*max(0; 5,38000000000000e+002-CDO)\*max(0; Y\_PER\_CDO\_S-5,29000000000000e+001) - 5,94818158861605e-003\*max(0; SED-7,00000000000000e-001)\*max(0; 2,72000000000000e+002-CDO\_S) + 2,85684358757798e-003\*max(0; CDO\_S-2,56000000000000e+002)\*max(0; Y\_PER\_CDO\_S-5,29000000000000e+001) - 2,68197780829575e-003\*max(0; 2,56000000000000e+002-CDO\_S)\*max(0; Y\_PER\_CDO\_S-5,29000000000000e+001) - 1,71960254140099e-004\*max(0; CDO-3,71000000000000e+002)\*max(0; 2,72000000000000e+002-CDO\_S) + 1,85444081927739e-004\*max(0; 3,71000000000000e+002-CDO)\*max(0; 2,72000000000000e+002-CDO\_S) - 4,17388972451093e-004\*max(0; CDO-3,38000000000000e+002)\*max(0; 5,75000000000000e+001-VSS) - 2,30581349545939e-001\*max(0; 7,30000000000000e+000-SED\_P) |
| Tableau 49: Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 93: les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 94: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 2 en valeur absolue** |

## **Réseaux de neurones (SANN)**

|  |
| --- |
| Tableau 50: Summary of active networks |

|  |  |
| --- | --- |
| Figure 95:les valeurs Prédites modèle 1 avec les valeurs Observes  **R²=0,98999887**  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 96: les valeurs Prédites modèle 2 avec les valeurs Observes  **R²=0,98926881**  **On remarque que toutes des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

**D’après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le premier réseau (R²) meilleur par rapport au deuxième**

|  |  |
| --- | --- |
| Figure 97:les valeurs Prédites avec les valeurs Observes | **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées) pour les deux modèles** |
| Figure 98:Analyse des résidus avec les valeurs Prédites modèle 1  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,106 sur 106 ont des résidus inférieurs à 0,04 en valeur absolue** | Figure 99:Analyse des résidus avec les valeurs Prédites modèle 2  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est bon. sur les variables prédites,106 sur 106 ont des résidus inférieurs à 0,8 en valeur absolue** |

# **VAR42 : Y\_GPERF\_SS sur l’ensemble test**

## **la méthode de Régression avec Backward Stepwise**

|  |  |
| --- | --- |
| **Prediction equation for**: **Y\_GPER\_SS** = 3,4753762005+0,0241986536467\*SS+0,00662112947999\*COND-0,0302303526268\*"SS\_P"-0,00604773853017\*"COND\_P"+5,16551507678\*"pH\_S"-0,0508129882773\*"BDO\_S"+0,0632563855931\*"SS\_S"+0,369130029229\*"Y\_PER\_SS\_P"-0,167648348374\*"Y\_PER\_SED\_P"+0,415607490903\*"Y\_PER\_BDO\_S"+0,0934081531343\*"Y\_PER\_CDO\_S" | |
| Tableau 51:Tests d'importance univariés pour Y\_GPER\_SS    Figure 100:Pareto Chart of t-Values for Coefficients  **D’après le tableau (p-valeu) on remarque que la méthode Backward Stepwise a éliminé le problème des variables non significatifs qui apparaitre dans notre modèle** | |
| Tableau 52:Paramètre Estimâtes et VIF   |  |  |  | | --- | --- | --- | |  | **Y\_GPER\_SS**  **paramétre** | **VIF** | | **Intercept** | **3,475376** |  | | **SS** | **0,024199** | **18,832760** | | **COND** | **0,006621** | **10,705649** | | **SS\_P** | **-0,030230** | **26,500826** | | **COND\_P** | **-0,006048** | **10,868229** | | **pH\_S** | **5,165515** | **1,442821** | | **BDO\_S** | **-0,050813** | **1,482337** | | **SS\_S** | **0,063256** | **2,024241** | | **Y\_PER\_SS\_P** | **0,369130** | **5,001749** | | **Y\_PER\_SED\_P** | **-0,167648** | **2,379668** | | **Y\_PER\_BDO\_S** | **0,415607** | **1,847746** | | **Y\_PER\_CDO\_S** | **0,093408** | **1,798900** |   **D’après le tableau on remarque qu’il y a le problème de multicolinéarité** | Tableau 53:R,Radj, SS Model et SS residual    **D’après le coefficient de détermination > 𝟎, 𝟓 (𝑹=0,86 > 𝟎, 𝟕), on remarque que on a une bonne corrélation entre Xi et Y. et on remarque que la somme de carrés résiduelle presque 35%, la partie de la variabilité de Xi qui n’est pas expliqué par notre modèle. Ce que qui montre que notre modèles est moyen bon** |
| Figure 101:les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |
| Figure 102: résidus sur échelle de probabilité gaussienne  **On remarque que notre modèle suit la loi normale** | Figure 103:Analyse des résidus avec les valeurs Observes |
| **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est moyen bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 10 en valeur absolue**  Figure 104: Analyse des résidus avec les valeurs Prédites | |

## **Modèle de Régression MARS**

0 cases with missing data were found.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MARSplines Results:

Dependent: Y\_GPER\_SS

Independents: FL, ZN, pH, BDO, CDO, SS, VSS, SED, COND, pH\_P, SS\_P, VSS\_P, SED\_P, COND\_P, pH\_S, BDO\_S, ...

Number of terms = 8

Number of basis functions = 11

Order of interactions = 2

Penalty = 2,000000

Threshold = 0,000500

GCV error = 12,940148

Prune = Yes

**Après la régression MARS, les variables retenues dans le modèle (6 sur 25) sont ZN, SS, VSS, pH\_S, BDO\_S, Y\_PER\_BDO\_S.**

|  |  |
| --- | --- |
| Tableau 54: Number of times each predictor is referenced    **Ce tableau nous donne le nombre de fois que la variable est présente dans l’expression de l’équation prédictive.**  **D’après le tableau le variable SS (4fois), est le plus référencée. Et les significative.** | Tableau 55:Regression statistics    **D’après le tableau on remarque que notre modèle prédit est moyen bon (R2 = 77,5% et R2ajusté = 75,6%), les moyennes des valeurs observées et prédites sont égales et leurs écarts-types proches (6,54 et 5,76)** |

|  |
| --- |
| **L’équation prédictive du modèle avec les variables indépendantes retenues**  NOTE: The following model should be used directly, with categorical variables being coded 0, 1.  **Y\_GPER\_SS** = 8,76698958820651e+001 - 4,99629898952080e-001\*max(0; 7,87000000000000e+001-Y\_PER\_BDO\_S) - 4,11892626536619e-002\*max(0; 2,82000000000000e+002-SS) + 1,00394574081919e-001\*max(0; 1,63000000000000e+002-BDO\_S) + 1,16150852239506e+000\*max(0; pH\_S-7,40000000000000e+000)\*max(0; Y\_PER\_BDO\_S-7,87000000000000e+001) - 2,38729589856534e-003\*max(0; 2,82000000000000e+002-SS)\*max(0; VSS-5,75000000000000e+001) - 5,40746797902832e-003\*max(0; 2,82000000000000e+002-SS)\*max(0; 5,75000000000000e+001-VSS) + 1,41398410262425e-002\*max(0; ZN-2,60000000000000e+000)\*max(0; 2,82000000000000e+002-SS) |
| Tableau 56: Model coefficients    **Ce tableau nous donne les noeuds employés dans le modèle et les coefficients** |

|  |  |
| --- | --- |
| Figure 105:les valeurs Prédites avec les valeurs Observes  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 106: Analyse des résidus avec les valeurs Prédites  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon. sur les variables prédites,105 sur 106 ont des résidus inférieurs à 8 en valeur absolue** |

## **Réseaux de neurones (SANN )**

**Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs**

|  |
| --- |
| Tableau 57: Summary of active networks |

|  |  |
| --- | --- |
| Figure 107:les valeurs Prédites modèle 1 avec les valeurs Observes  **R²= 0,69082073**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** | Figure 108: les valeurs Prédites modèle 2 avec les valeurs Observes  **R²= 0,6312257**  **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées)** |

**D’après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le premier réseau (R²) meilleur par rapport au deuxième**

|  |  |
| --- | --- |
| Figure 109:les valeurs Prédites avec les valeurs Observes | **On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l’équation prédictive linéaire du modèle avec les valeurs observées) pour les deux modèles** |
| Figure 110:Analyse des résidus avec les valeurs Prédites modèle 1  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites,103 sur 106 ont des résidus inférieurs à 8 en valeur absolue** | Figure 111: Analyse des résidus avec les valeurs Prédites modèle 2  **D’après le graphe de l’analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu est globalement bon. sur les variables prédites, 103 sur 106 ont des résidus inférieurs à 8 en valeur absolue** |

# **Comparer la performance de nos modèles sur l’ensemble train :**

## **Var40 :Y\_GPER\_BDO**

Tableau 58: comparaison des modèles sur l'ensemble train VAR40

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRO | **73,19%** | **71,50%** | Bon | **Moyenne**  **sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **signification, multicolonéarité, points aberrantes.** |
| MRB | **72,43%** | **71,83%** | **Bon** | **Moyenne**  **sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **points aberrantes** |
| MRB sans points aberrantes | **89,99%** | **89,97** | **Très bon** | **Bon**  **sur les variables prédites,419 sur 420 ont des résidus inférieurs à 10 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **Risque de perdre des données importantes** |
| MARS | **89,34%** | **88,94%** | **Très bon** | **Excellent**  **sur les variables prédites,418 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones | **79,64%** |  | **Bon** | **Très bonne**  **sur les variables prédites,418 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau et les remarques on voit que MRO (Multiple régression) et MRB (MR avec *Backward Stepwise*) sont valide pour prédite Y\_GPER\_BDO mais contient beaucoup des problémes(signification, multicolonéarité, points aberrantes), et lorsque nous avons essayé de résoudre ces problémes avec MRB sans points aberrantes nous avons trouvé un trés bon modèle, mais avec ce modèle on risque de perdre des données importantes.**

**avec MARS et SANN nous avons trouvé des bons modèles, et on sait que la méthode SANN basée sur les données observé (problème overfitting (sur apprentissage)), donc on conclure le modèle de MARS est le meilleur choix.**

**-pour prédite Y\_GPER\_BDO sur l’ensemble train on peut utiliser MARS et SANN**

## **Var41 :Y\_GPER\_CDO**

Tableau 59: comparaison des modèles sur l'ensemble train VAR41

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRO | **85,52%** | **84,99%** | **Bon** | **Moyenne**  **sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **signification, multicolonéarité.** |
| MRB | **85,24%** | **84,99%** | **Bon** | **Bon**  **sur les variables prédites,419 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** |  |
| MARS | **94,49%** | **94,27%** | **Excellent** | **Excellent**  **sur les variables prédites,420 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones | **94,07%** |  | **Excellent** | **Excellent**  **sur les variables prédites,419 sur 421 ont des résidus inférieurs à 8 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau on remarque que avec MRO on a un bon modèle mais contient des problémes (signification, multicolonéarité), et avec MRB on a reussi a résoudre ces problémes, et on a obtenu des execellents modèles avec MARS et SANN , mais on remarque que le modèle de MARS est le meilleur choix.**

**-pour prédite Y\_GPER\_CDO sur l’ensemble train on peut utiliser MRB, MARS et SANN**

## **Var42 :Y\_GPER\_SS**

Tableau 60: comparaison des modèles sur l'ensemble train VAR42

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRO | **39,98%** | **36,18%** | **Mauvais** | **Mauvais**  **sur les variables prédites,417 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **Corrélation, signification, multicolonéarité, points aberrantes.** |
| MRB | **38,73%** | **37,39%** | **Mauvais** | **Mauvais**  **sur les variables prédites,417 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **Corrélation, multicolonéarité** |
| MRB sans points aberrantes | **63,99%** | **63,01%** | **Bon** | **Bon**  **sur les variables prédites,416 sur 417 ont des résidus inférieurs à 15 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **Risque de perdre des données importantes** |
| MARS | **59,68%** | **58,29%** | **Bon** | **Bon**  **sur les variables prédites,420 sur 421 ont des résidus inférieurs à 20 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones (SANN) | **62,9%** |  | **Bon** | **Bon**  **sur les variables prédites,421 sur 421 ont des résidus inférieurs à 10 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau et les remarques on voit que MRO (Multiple régression) et MRB (MR avec *Backward Stepwise*) ne sont pas valide(R<0,7) pour prédite Y\_GPER\_SS, et lorsque nous avons élimini les points aberrantes nous avons trouvé un bon modèle, mais avec ce modèle on risque de perdre des données importantes.**

**avec MARS et SANN nous avons trouvé des bons modèles, d’après R2 on remarque que SANN est le meilleur choix mais on sait que la méthode SANN basée sur les données observé (problème overfitting (sur apprentissage)), donc on conclure le modèle de MARS est le meilleur choix.**

**-pour prédite Y\_GPER\_SS sur l’ensemble train on peut utiliser MARS et SANN**

# **Comparer la performance de vos modèles sur l’ensemble test**

## **Var40 :Y\_GPER\_BDO**

Tableau 61: comparaison des modèles sur l'ensemble test VAR40

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRB | **93,44%** | **93,04%** | **Excellent** | **Excellent**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 6 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** |  |
| MARS | **99,44%** | **99,36%** | **Excellent** | **Excellent**  **sur les variables prédites,106 sur 106 ont des résidus inférieurs à 2,5 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones (SANN) | **98,7%** |  | **Excellent** | **Excellent**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 1,5 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau on remarque qu’avec les trois modèles MRB, MARS, SANN nous avons des execellents modèles, on remarque que le modèle de MARS est le meilleur choix.**

**-Pour prédite Y\_GPER\_BDO sur l’ensemble test on peut utiliser MRB, MARS et SANN**

## **Var41 :Y\_GPER\_CDO**

Tableau 62: comparaison des modèles sur l'ensemble test VAR41

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRB | **93,58%** | **93,05%** | **Excellent** | **Excellent**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 8 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** |  |
| MARS | **99,49%** | **99,39%** | **Excellent** | **Excellent**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 2 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones | **98,99%** |  | **Excellent** | **Excellent**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 0,04 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau on remarque qu’avec les trois modèles MRB, MARS, SANN nous avons des execellents modèles, on remarque que le modèle de MARS est le meilleur choix.**

**-Pour prédite Y\_GPER\_CDO sur l’ensemble test on peut utiliser MRB, MARS et SANN**

## **Var42 :Y\_GPER\_SS**

Tableau 63: comparaison des modèles sur l'ensemble test VAR42

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modèle | R2 | Radj 2 | Prédiction | Analyse des résidus | Complexité du modèle | Problème |
| MRB | **74,04%** | **71%** | **Bon** | **Bon**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 10 en valeur absolue** | **Facile (les étapes et la compréhension de résultats de modèle)** | **multicolonéarité** |
| MARS | **77,52%** | **75,67%** | **Bon** | **Bon**  **sur les variables prédites,105 sur 106 ont des résidus inférieurs à 8 en valeur absolue** | **Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)** | **Méthode un peu complexe** |
| Réseaux de neurones | **69,08%** |  | **Bon** | **Bon**  **sur les variables prédites,103 sur 106 ont des résidus inférieurs à 8 en valeur absolue** | **Étapes moyenne et la compréhension de résultats est difficile** | **-Méthode un peu complexe**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**D’après le tableau on remarque qu’avec les trois modèles MRB, MARS, SANN nous avons des bons modèles, on remarque que le modèle de MARS est le meilleur choix (R2=77,52%).**

**-Pour prédite Y\_GPER\_BDO sur l’ensemble test on peut utiliser MRB, MARS et SANN**

Tableau 64: choix des modèles pour le système

|  |  |  |
| --- | --- | --- |
|  | l’ensemble train | l’ensemble test |
| Y\_GPER\_BDO | **MARS et SANN** | **MRB, MARS et SANN** |
| Y\_GPER\_CDO | **MRB, MARS et SANN** | **MRB, MARS et SANN** |
| Y\_GPER\_SS | **MARS et SANN** | **MRB, MARS et SANN** |

**D’après les deux comparaison précédent et le tebleau on conclure que pour prédit notre systéme (Y\_GPER\_BDO, Y\_GPER\_CDO, Y\_GPER\_SS) on peut utilisé MARS et SANN. Et on remarque que MARS est le meilleur choix**

**On remarque que le choix de méthode pour la modélisation dépend de quantité des données (ensemble train et ensemble test), comme on a vu lorsque on a utilisé MRO et MRB pour modéliser l’ensemble train on n’a pas obtenu des bon résultat, par contre avec l’ensemble test on a eu des bon résultat.**

**Nous avons vu aussi qu’avec les deux méthodes MARS et SANN on a eu des très bons résultats pour les deux ensembles.**

**Et on a trouvé que la meilleure méthode permit nos méthodes est MARS.**

# **D) une conclusion sommaire**

Tableau 65: les forces et faiblesses de nos méthodes

|  |  |  |
| --- | --- | --- |
|  | forces | faiblesses |
| MRO | **-les résultats donnent directement l’importance relative des variables d’entrée sur le modèle**  **-Facile a applique** | **-problème de signification**  **-problème de multicolonéarité**  **-résultat peut etre biaisés (résultats aberrants, manque de robustesse..) pour une très grande quantité des données (manque des données…).**  **-Incapable de découvrir la structure locale des données** |
| MRB | **-résous le problème de multicolinéarité (élimine les variables explicatives les moins importantes dans les données)**  **-les résultats donnent directement l’importance relative des variables d’entrée sur le modèle a partir de leurs effets significatifs**  **-le meilleur choix par rapport MRO et MRF**  **-Facile a applique** | **-Pouvoir prédictif et analyse des résidus faible par rapport MARS et Réseaux de neurones**  **-résultat peut etre biaisés (résultats aberrants, manque de robustesse..) pour une très grande quantité des données (manque des données…).**  **-Incapable de découvrir la structure locale des données** |
| MARS | **-Facilement a utilisé pour exploiter une très grande quantité des données complexe**  **-Capable de découvrir la structure locale des données**  **-Pouvoir prédictif et analyse des résidus très bon par rapport à la régression multiple** | **-les étapes et la compréhension de résultats de modèle sont un peu difficile par rapport MRB**  **-méthode un peu complexe par rapport MRB** |
| Réseaux de neurones | **-Facilement a utilisé pour exploiter une très grande quantité des données complexe**  **-Pouvoir prédictif et analyse des résidus très bon par rapport à la régression multiple** | **-les étapes et la compréhension de résultats de modèle sont difficile par rapport MRB et MARS**  **-méthode un peu complexe par rapport MRB**  **-overfitting (sur apprentissage), méthode basée sur les données observé** |

**Dans la question C nous avons vu que le choix de méthode pour la modélisation dépend de quantité des données (ensemble train et ensemble test), comme on a vu lorsque on a utilisé MRO et MRB pour modéliser l’ensemble train on n’a pas obtenu des bons résultats, par contre avec l’ensemble test on a eu des bon résultat.**

**Nous avons vu aussi qu’avec les deux méthodes MARS et SANN on a eu des très bons résultats pour les deux ensembles.**

**Et on a trouvé que la meilleure méthode permit nos méthodes est MARS.**

**Le processus de modélisation statistique à l’aide de modèles de régression incluant la méthode les réseaux de neurones serait définir le problème et identifier et préparer les données, et nous permit de construire, tester et évaluer le modèle en choisissant la technique la plus optimale et trouver le meilleur réseau. Et pour aide au traitement la méthode de MARS nous permit à exploiter une très grande quantité des données complexe et utilise l’algorithme itératif adaptif utile pour aider le réseau à effectuer le traitement.**

**E) Comparer les données d’entrée de 1990 et celles de 1991.**

**Pour faire la comparaison entre les données on utilise la méthode t-test (t-test for independent samples by groups)**

|  |
| --- |
| Tableau 66: T-tests    **D’aprés le tableau on remarque que :**  **La variable FL plus important dans les deux années par rapports les autres variables**  **D’après p-valeu FL, pH, CDO, VSS, COND sont les variables significatifs leur P<0,05**  **Variables FL , COND, BDO: on remarque que la moyenne pour le groupe 1990 supérieure a la moyenne pour le groupe 1991 et l’écart type pour le groupe 1990 supérieure a l’écart type pour le groupe 1991 qui normalement implique que la dispersion dans le groupe 1990 élevée par rapport le groupe 1991 mais le nombre de validation pour le groupe 1990 supérieure le groupe 1990 donc on peut rien dire.**  **Variables CDO: on remarque que la moyenne pour le groupe 1990 inférieure a la moyenne pour le groupe 1991 et l’écart type pour le groupe 1990 inférieure a l’écart type pour le groupe 1991 et aussi le nombre de validation pour le groupe 1990 supérieure le groupe 1990 qui implique que pour le variable CDO le groupe 1990 plus regroupé par contre le groupe 1991 plus dispersive.**  **Variable pH, ZN : on remarque que la moyenne pour le groupe 1990 proche a la moyenne pour le groupe 1991 et l’écart type pour le groupe 1990 proche a l’écart type pour le groupe 1991.**  **Variable SS : on remarque que la moyenne pour le groupe 1990 supérieure a la moyenne pour le groupe 1991 et l’écart type pour le groupe 1990 inférieure a l’écart type pour le groupe 1991 qui implique que pour le variable SS le groupe 1990 plus regroupé par contre le groupe 1991 plus dispersive.**  **Variable VSS : on remarque que la moyenne pour le groupe 1990 inférieure a la moyenne pour le groupe 1991 et l’écart type pour le groupe 1990 supérieure a l’écart type pour le groupe 1991 .** |

|  |  |
| --- | --- |
| Figure 112: Box & Whisker Plot: pH  **D’après le graphe on remarque que la mesure de pH dans l’année 1990 plus important par rapport l’année 1991** | Figure 113: Box & Whisker Plot: SS  **D’après le graphe on remarque que la mesure de SS plus regroupé dans l’année 1990 et plus dispersive dans l’année 1991** |
| Figure 114: Box & Whisker Plot: SED  **D’après le graphe on remarque que la mesure de VSS dans l’année 1991 plus important et plus dispersive par rapport l’année 1990 qui plus regroupé** | Figure 115: Box & Whisker Plot: VSS  **D’après le graphe on remarque que la mesure de VSS dans l’année 1991 plus important par rapport l’année 1990** |
| Figure 116: Box & Whisker Plot: COND  **D’après le graphe on remarque que la mesure de COND dans l’année 1990 plus important par rapport l’année 1991** | Figure 117: Box & Whisker Plot: BDO  **D’après le graphe on remarque que la mesure de BDO dans l’année 1990 plus important par rapport l’année 1991** |
| Figure 118: Box & Whisker Plot: CDO  **D’après le graphe on remarque que la mesure de CDO dans l’année 1991 plus important et plus dispersive par rapport l’année 1990 qui plus regroupé** | Figure 119: Box & Whisker Plot: ZN  **D’après le graphe on remarque que la mesure de ZN dans l’année 1991 plus important et plus dispersive par rapport l’année 1990 qui plus regroupé** |
| Figure 120: Box & Whisker Plot: FL  **D’après le graphe on remarque que la mesure de FL dans l’année 1990 plus important par rapport l’année 1991** |  |

**D’après les graphes précédents on remarque que dans l’année 1990 les mesures pH, COND, FL, BOD sont élevés et joue un rôle important par rapport l’année 1991, cependant les mesures SED, VSS, CDO, ZN, SS sont élevés et joue un rôle important par rapport l’année 1990 dans l’année 1991**

**Remarque : dans cette question on peut utilisé la méthode One-way ANOVA pour faire la comparaison**

**Question bonus :**

**Pour la Question bonus on choisi la régression RIDGE puisque : 1-permet d’estimer un modèle en présence de covariables fortement corrélées. 1-estimation dépendante de l’échalle des variables (centrer et réduire toutes les variables continues).**

**En générale**

**Nos données sont Quantitatif soit fixe ou Continu donc pour analyser les données nous utilisons :**

**-Modèle de Régression Ordinaire (MRO) (RÉGRESSION LINÉAIRE MULTIPLE)**

**Et comme nous avons trouvé le problème de multicolonéarité on peut utiliser**

**-la régression en composantes principales (ACP),**

**-la régression RIDGE,**

**-la méthode de sélection de variables**

**Et pour résoudre le problème de signification on peut utiliser**

**-Modèle de Régression avec Sélection pas à pas Avant (Forward Stepwise) (MRF)**

**-Modèle de Régression avec Sélection pas à pas Arrière (Backward Stepwise) (MRB)**

# **Conlusion générale :**

**Dans cette devoir nous avons étudié un cas réel de données ou nous avons appliqué les méthodes que nous avons faites dans notre cours, et nous avons trouvé des vrais problèmes de réalité.**

**Nous avons pris que le choix de modèle dépend de type et le rôle des données et aussi dépend de types de problème trouvé (l’influence de signification des données et la multicolonéarité et les points aberrantes). On a pris comment résoudre ces problèmes avec le choix de modèle. Nous avons aussi comparé l’influence du plusieurs modèles sur les données**

**Nous avons utilisé la représentation graphique des données qui nous donne une idée de la nature de relation entre les variables explicatives et les variables expliauées. Et aussi l’analyse des résidus qui est une composante essentielles pour la validation du modèlel. Et nous avons vu l’influence et l’impoetance de préparation des données brutes.**