IAE Nantes
Économie & Management
Pôle Sociétés
Nantes Université

Nantes Université

Master 1 - Econométrie et Statistiques, parcours Econométrie Appliquée

Modélisation dans le cadre de la performance thermique

Dossier réalisé par :
QUINTIN DE KERCADIO Pierre
CROCHET Florian

Année universitaire 2024 - 2025

Sommaire

I - Introduction

L'objectif de ce projet est de modéliser la consommation énergétique des bâtiments universitaires à l'aide de deux méthodes basées sur les variables latentes : la régression sur composantes principales, PCR, et la régression des moindres carrés partiels, PLS. Les données, issues de simulations sur 144 bâtiments des universités de UPENN et GT (Tian et al., 2015), incluent des caractéristiques géométriques, thermiques et d'usage.

Le projet vise à comparer la performance prédictive des deux approches, tout en identifiant les variables les plus influentes sur la consommation énergétique.

II - Exploration des données

Le jeu de données étudié comporte 33 variables dont une variable cible, la charge énergétique totale des bâtiments (EnergyLoad), que nous avons construite en additionnant les charges de chauffage (HeatTotal) et de refroidissement (CoolTotal). Les prédicteurs (X) incluent des variables géométriques (surface, hauteur), thermiques (coefficients de transmission thermique, surfaces vitrées), et d'usage (densité d'occupation, équipements, éclairage, etc.).

La distribution de EnergyLoad[1] révèle une forte asymétrie à droite, marquée par un regroupement des observations vers les faibles valeurs et quelques extrêmes. Cette distribution non gaussienne peut nuire aux performances des modèles linéaires, donc pour y remédier, nous avons appliqué une transformation en racine carrée, avec pour but de réduire la skewness et stabiliser la variance. Le second graphique montre que la variable transformée (sqrt_EnergyLoad) présente une distribution beaucoup plus symétrique. Ce constat est appuyé par le test de Shapiro-Wilk, qui affiche une amélioration du coefficient W (de 0,62 à 0,93) après transformation.

L'objectif de transformer la variable dépendante (Y) en racine carrée est d'améliorer la performance du modèle, en stabilisant la variance des erreurs et en réduisant l'impact des valeurs extrêmes. Le graphique des résidus[2] confirme l'intérêt de cette transformation : dans le modèle sans transformation, les résidus sont très dispersés, avec des valeurs extrêmes (−61,2 et 158), suggérant une hétéroscédasticité potentielle et que certaines observations pourraient avoir une influence disproportionnée sur le modèle. En revanche, après transformation de Y en racine carrée, les résidus sont recentrés et beaucoup moins dispersés (entre −5 et 5), ce qui témoigne d'une réduction de l'étendue des résidus et d'une meilleure homogénéité, rendant les erreurs plus constantes et moins influencées par les valeurs extrêmes. Cette amélioration de la distribution des erreurs indique que la transformation réduit l'influence des valeurs extrêmes et améliore l'homogénéité des erreurs, ce qui pourrait potentiellement améliorer la performance du modèle. La transformation semble donc justifiée au regard de la distribution des erreurs. Il conviendra désormais d'évaluer si cette amélioration se traduit réellement par de meilleures performances du modèle.

[1] Chunk 8 : distribution de Y non transformée (EnergyLoad) et Y transformée (sqrt_EnergyLoad)
[2] Chunk 41 : graphique des résidus

De plus, l'analyse des prédicteurs révèle d'importantes variations d'échelle (ex. : zone_area varie de quelques centaines à plus de 16 000), ce qui justifie leur centrage et réduction avant toute modélisation. L'histogramme de chaque variable[3] montrent également la présence de valeurs extrêmes, ce qui pourrait impacter les erreurs quadratiques.

Enfin, la matrice de corrélation[4] et le VIF[5] mettent en évidence une multicolinéarité importante entre plusieurs groupes de variables : les surfaces orientées (Nord, Sud...), les surfaces vitrées, ou encore les variables d'usage (occupants, équipements, éclairage). Cette forte corrélation entre les variables explicatives justifie le recours à des méthodes de modélisation basées sur des composantes latentes, telles que la PCR ou la PLS, qui sont plus robustes que la régression classique face à la multicolinéarité.

Par la suite, une séparation des données a été réalisée entre un jeu d'apprentissage (UPENN), utilisé pour l'entraînement des modèles et la validation croisée, et un jeu de test (GT), destiné à évaluer la capacité de généralisation des modèles sur des observations totalement nouvelles.

III - Principal Components Regression (PCR)

La méthode de régression sur composantes principales (PCR) a été appliquée pour modéliser la charge énergétique totale des bâtiments, après transformation de la variable réponse par racine carrée. Cette approche nous permet de contourner les problèmes de multicolinéarité en utilisant des composantes principales orthogonales, tout en réduisant la dimensionnalité du jeu de données et en conservant l'information la plus significative contenue dans les variables explicatives.

Nous avons pu observer l'évolution de l'erreur quadratique moyenne, le RMSE, sur les jeux d'apprentissage et en validation croisée. Le graphique[6] montre une diminution progressive du RMSE CV, qui se stabilise à partir de la vingtième composante, seuil retenu comme optimal selon la règle du « one sigma »[7]. De plus, les performances du modèle sont satisfaisantes sur les données internes : le R2 atteint 77,8 % en apprentissage et 64,2 % en validation croisée[8], ce qui signifie une capacité d'ajustement correcte et un risque modéré de surapprentissage.

Cependant, le passage au jeu test révèle une forte chute de performance. En effet, le R2 test[9] chute à −50,9 %, indiquant que le modèle prédit moins bien que la simple moyenne des observations. Le RMSEP test[10] s'élève à 8,90, soit une erreur quasiment équivalente à celle obtenue avec une prédiction constante. Ce phénomène s'explique par le fait que la PCR sélectionne les composantes uniquement en fonction de la variance des prédicteurs, sans tenir compte de leur relation avec la variable à expliquer. Cette stratégie peut conduire à retenir certaines directions

[3] Chunk 14 : histogramme de chaque variable
[4] Chunk 16 : matrice de corrélation
[5] Chunk 32 : VIF des variables explicatives
[6] Chunk 63 : valeurs et graphique du RMSE (apprentissage et validation croisée)
[7] Chunk 64 : critère « one sigma » pour le choix du nombre de composantes
[8] Chunk 62 : valeurs et graphique du R^2 (apprentissage et validation croisée)
[9] Chunk 66 : valeurs du R^2 sur le jeu test
[10] Chunk 67 : valeurs du RMSE sur le jeu test

principales qui résument bien X, mais qui sont peu informatives pour Y, ce qui limite la capacité de généralisation du modèle.

Lorsque l'on applique la PCR directement sur la variable EnergyLoad non transformée, les performances sont moins bonnes[11], en particulier en apprentissage (R^2 = 45,5 % et RMSE = 25,1) ainsi qu'en validation croisée (R^2 CV = 23,6 % et RMSE CV = 29,7), où la qualité de l'ajustement du modèle diminue et la dispersion des erreurs de prédiction augmente. Le RMSEP est d'autant plus élevé (135,8) et le R^2 test reste négatif (−18,9 %). Ces résultats indiquent que la transformation en racine carrée améliore l'ajustement du modèle aux données d'entraînement.

IV - Partial Least Squares Regression (PLS)

La méthode des moindres carrés partiels (PLS) a été utilisée pour modéliser la charge énergétique totale des bâtiments, à partir de la variable transformée par racine carrée. Cette approche vise à construire des composantes latentes en maximisant la covariance entre les variables explicatives et la variable à prédire, ce qui permet de sélectionner des directions pertinentes à la fois du point de vue de la variance des prédicteurs et de leur pouvoir explicatif sur la variable cible Y.

La PLS est particulièrement adaptée à des jeux de données présentant une forte multicolinéarité, comme c'est le cas ici, puisque de nombreuses variables explicatives sont corrélées entre elles (notamment les surfaces orientées, les variables thermiques et les usages internes). En intégrant directement l'information de Y dans la construction des composantes, la méthode permet de capter plus rapidement les directions les plus informatives pour la prédiction.

L'évolution de l'erreur quadratique moyenne a été examinée sur les jeux d'apprentissage et en validation croisée. Le graphique associé[12] montre une nette diminution du RMSE jusqu'à la troisième composante, après quoi la courbe se stabilise. De plus, le critère « one sigma » fixe le nombre optimal de composantes[13] à 3.

Les performances du modèle sont très satisfaisantes en apprentissage, avec un R^2 de 80,1 %, ce qui indique une bonne capacité du modèle à expliquer la variabilité de sqrt_EnergyLoad sur les données UPENN. En validation croisée, le R^2 atteint 67,0 %, traduisant une généralisation correcte, et un RMSE CV de 1,61, inférieur à celui obtenu avec la PCR. Ces résultats[14] montrent que la PLS, en orientant les composantes principales vers la prédiction de Y, parvient à construire un modèle plus efficace avec un nombre réduit de composantes.

Cependant, comme observé pour la PCR, les performances chutent significativement sur le jeu test. Le R^2 devient fortement négatif (−49,6 %), ce qui signifie que le modèle est moins performant que la simple prédiction par la moyenne de la variable à expliquer. Le RMSEP sur le jeu test atteint 8,78, soit une erreur du même ordre que celle observée pour la PCR. Ces résultats suggèrent que, bien que le modèle soit bien ajusté aux données d'entraînement, il souffre d'un

[11] Chunk 70 : Résultats de la base avec EnergyLoad et de celle avec sqrt_EnergyLoad
[12] Chunk 90 : valeurs et graphique du RMSE (apprentissage et validation croisée)
[13] Chunk 91 : critère « one sigma » pour le choix du nombre de composantes
[14] Chunk 97 : valeurs des indicateurs de performance R^2 et RMSE

manque de robustesse dès qu'il est confronté à une structure de données différente, ici représentée par les bâtiments d'un autre campus.

Lorsque l'on applique la PLS directement sur la variable EnergyLoad non transformée, les performances sont globalement moins bonnes que celles obtenues avec la transformation en racine carrée. En apprentissage, le R^2 atteint seulement 56,8 %, et chute à 32,0 % en validation croisée, ce qui traduit une capacité d'ajustement plus limitée. De plus, un RMSE CV plus élevé (28,03), un R^2 en test négatif (−19,1 %) et un RMSEP en test de 136,63 indiquent de nouveau une très mauvaise généralisation. Ces résultats[15] confirment que la transformation en racine carrée contribue à stabiliser le modèle, en réduisant l'influence des valeurs extrêmes et en améliorant l'homogénéité des erreurs[16], ce qui facilite l'identification de composantes latentes pertinentes et améliore potentiellement la performance prédictive.

V - Comparaison des modèles

La comparaison entre les deux méthodes montre que la PLS est plus efficace que la PCR. Cela s'explique par le fait que la PLS cherche à maximiser la covariance entre les composantes construites à partir de X et la variable Y. Par conséquent, les directions retenues par la PLS sont orientées vers la prédiction, alors que la PCR se contente de maximiser la variance expliquée de X, sans tenir compte de Y. Dans un contexte où de fortes corrélations existent entre les variables explicatives, comme c'est le cas ici, la PLS est donc plus apte à extraire rapidement l'information utile à la prédiction, et ce avec un nombre de composantes réduit.

Les résultats[17] révèlent ainsi que le modèle PLS appliqué à la variable transformée par la racine carrée atteint un R^2 de 80,1 % en apprentissage et 67,0 % en validation croisée, contre respectivement 77,8 % et 64,2 % pour la PCR, avec seulement 3 composantes contre 20 pour la PCR. De plus, les erreurs de prédiction sont également plus faibles avec la PLS, avec un RMSE de 1,25 en apprentissage et de 1,61 en validation croisée, contre 1,32 et 1,68 pour la PCR. Ces résultats traduisent une plus grande efficacité de la PLS à capter la structure informative du lien entre X et Y, ce qui lui permet de conserver une bonne capacité de prédiction sur les données d'entraînement. Cependant, avec des R^2 en test négatifs et des RMSEP nettement plus élevés que sur les données d'entraînement, les modèles présentent une faible validité externe, c'est-à-dire qu'ils peinent à généraliser leurs performances à de nouvelles données, comme celles provenant des bâtiments d'un autre campus.

Pour interpréter notre modèle PLS avec Y transformée, nous avons utilisé l'analyse des scores VIP[18] qui permet d'identifier les variables contribuant le plus à la prédiction de la charge énergétique. Parmi les variables les plus influentes (VIP > 1), plusieurs groupes se distinguent.

Tout d'abord, plusieurs variables décrivent le bâtiment, avec notamment zone_area (surface totale) et bldg_height (hauteur), deux variables fortement contributives. Un bâtiment plus vaste ou

[15] Chunk 98 : résultats de la base avec EnergyLoad et de celle avec sqrt_EnergyLoad
[16] Chunk 41 : graphique des résidus
[17] Chunk 99 : résultats de l'ensemble des modèles
[18] Chunk 100 : VIP du modèle PLS avec Y transformée

plus haut implique un volume à chauffer ou refroidir plus important, ce qui accroît logiquement la consommation énergétique. On retrouve également une influence marquée des surfaces d'ouvertures orientées (op_S_area, op_E_area, op_N_area, op_W_area), qui constituent autant de zones d'échange thermique avec l'extérieur. La variable gl_S_area (surface vitrée exposée au sud) complète ce groupe : son rôle est directement lié aux apports solaires reçus en façade sud, susceptibles d'influer sur les besoins de chauffage en hiver ou de climatisation en été.

Du côté des caractéristiques thermiques, trois variables ressortent très nettement : op_uValue, roof_op_uValue et gl_U_value. Ces coefficients traduisent le niveau d'isolation des ouvertures, du toit et des vitrages. Un coefficient U élevé indique une mauvaise isolation et donc des pertes thermiques plus importantes. Ces variables se révèlent essentielles pour expliquer la variabilité de la consommation énergétique. À cela s'ajoute la surface d'ouverture en toiture (roof_op_area), également impliquée dans les échanges thermiques, et gl_solar_trans, qui mesure la transmission solaire des vitrages.

Enfin, les gains de chaleurs internes jouent aussi un rôle majeur. La variable totalocc (nombre d'occupants) affecte directement les apports internes de chaleur, les besoins en ventilation, et donc la consommation. On retrouve également deux variables directement associées aux flux thermiques internes : heat_flow_rate (chauffage) et cool_flow_rate (climatisation), qui traduisent les besoins énergétiques effectifs. Enfin, la variable chair_per_occ, qui peut être vue comme un indicateur indirect d'intensité d'utilisation (densité d'équipement), suggère un lien entre la fréquentation des espaces et leur demande énergétique. En complément, gl_E_area et gl_N_area, bien qu'en-dessous du seuil 1, présentent des scores VIP autour de 0.9, ce qui justifie leur présence dans l'analyse. Elles traduisent des influences directionnelles modérées sur les pertes et gains thermiques.

VI - Conclusion

En conclusion, la transformation en racine carrée de la variable cible a significativement amélioré la stabilité des modèles et leur qualité d'ajustement, par rapport aux versions sans transformation, où les erreurs étaient plus dispersées et les performances inférieures. Parmi les méthodes évaluées, la PLS s'est révélée la plus efficace, avec un modèle optimal reposant sur uniquement 3 composantes latentes. Ce modèle surpasse nettement la PCR en termes de pouvoir prédictif et de parcimonie. Toutefois, bien que la PLS ait été plus performante en apprentissage, les deux modèles ont montré une faible validité externe, soulignant la difficulté de généraliser les résultats à de nouvelles données et la nécessité d'améliorer leur robustesse. Enfin, l'analyse des scores VIP a permis d'identifier 17 variables principales influençant la consommation énergétique des bâtiments, principalement liées à leur géométrie, à l'isolation thermique et aux usages internes. Ces résultats confirment l'intérêt des méthodes à variables latentes pour modéliser des systèmes complexes en présence de multicolinéarité.