ISEP
IF.1205 – Science et Traitement des données
Mai 2023

PROJET DE SCIENCE DES DONNEES

Analyse en Composantes Principales (ACP) et Régression Linéaire

I. Instructions générales

Le but de ce projet est d'appliquer les méthodes d'analyse en composantes principales et de régression linéaire sur des données réelles. Le projet sera réalisé par équipe de **trois étudiants.** Chaque équipe devra présenter son travail au cours d'une soutenance orale qui aura lieu le **mardi 30 mai prochain**. Vous pouvez travailler en R ou en Python. Les instructions concernant la soutenance orale et l'évaluation sont précisées ci-dessous.

a) Instructions concernant la soutenance orale

La soutenance orale durera environ 15 minutes par groupe, et se décomposera en 10 minutes de présentation et 5 minutes de questions. Les diapositives de la présentation devront contenir les éléments suivants :
- Une page de couverture contenant le prénom, nom de famille, et le numéro d'identification étudiant de tous les membres de l'équipe.
- Un sommaire.
- Une courte introduction.
- Le corps du document (résultats, figures, tables, interprétations, commentaires, ou tout autre élément qui permette de répondre aux questions). Les réponses aux questions posées dans le sujet doivent être clairement indiquées dans cette partie. Si besoin, vous pouvez utiliser jusqu'à trois nombres significatifs dans vos applications numériques.
- La conclusion.
- Les références.

Votre code R ou Python ne doit pas être inclus dans la présentation. Cependant, vous devez avoir votre code sous la main au moment de la soutenance, pour répondre à toute question éventuelle à ce sujet.
Un seul membre de votre équipe devra déposer votre fichier de présentation au format pdf sur Moodle le **29 mai au plus tard**. Un dépôt Moodle sera créé dans ce but pour chaque groupe de TD (G7, G8, G9 et G10). Le nom du fichier déposé devra respecter la forme suivante :
NomEtudiant1_NomEtudiant2_NomEtudiant3.pdf

b) Instructions concernant l'évaluation

La soutenance orale sera divisée en deux parties : 10 minutes de présentation orale et 5 minutes de questions. L'évaluation sera à la fois collective, notamment pour la qualité et le contenu global de la présentation, mais aussi individuelle.

Chaque étudiant sera donc évalué sur ses interventions au cours des deux parties. Une attention particulière sera portée à la qualité des réponses en termes d'analyse.

II. Analyse de données

a) Le jeu de données

Les programmes de sciences participatives proposent à des citoyens de collecter bénévolement des données, généralement sur une plateforme, afin qu'elles soient ultérieurement analysées par des chercheurs. Ils sont de plus en plus répandus, notamment pour faire l'état des lieux de la biodiversité. Par exemple, certains consistent à suivre l'évolution de populations d'oiseaux afin de déterminer les espèces en déclin, ou au contraire celles qui résistent.

Dans ce contexte, le but de ce projet est d'analyser les résultats du recensement participatif sur une année des espèces d'oiseaux observés dans un parc parisien. On s'intéressera plus particulièrement au suivi des perruches à collier, considérées comme invasives en France depuis 2018, et à leur impact sur les autres populations d'oiseaux.

Le nombre d'oiseaux observé par mois et par espèce au cours d'une année dans un même parc parisien est recensé dans le fichier **data.csv**.

b) Analyse préliminaire : statistiques descriptives

Importez le jeu de données data.csv. Afin de vous familiariser avec les données, répondez aux questions suivantes :

1. Combien d'individus, toute espèce confondue, et d'espèces différentes ont été observés dans le parc ? Quelle espèce a été la plus observée dans l'année ? Quelle espèce a été la moins observée dans l'année ? Pour chacune de ces deux espèces, combien d'individus ont été comptabilisés ?
2. Certaines valeurs sont égales à 0 dans votre fichier. Interprétez. Combien de mois et d'espèces différentes sont concernés ? **Supprimez ces espèces. Elles ne seront pas prises en compte dans la suite de notre analyse.**
3. Calculez la variance du nombre d'oiseaux observés pour chaque espèce restante. Commentez. Quelle est l'espèce dont le nombre d'observations par mois varie le plus ? Quelle est l'espèce dont le nombre d'observations par mois varie le moins ?

On décide de s'intéresser plus particulièrement aux perruches à collier.

4. Calculez la moyenne, la médiane, le minimum, le maximum et l'écart-type du nombre de perruches observées. Affichez l'histogramme de ces observations. Commentez.
5. Affichez l'évolution du nombre moyen d'individus observés par mois, toute espèce confondue en ne prenant pas en compte les perruches à collier. Superposez sur cette même figure l'évolution du nombre de perruches à collier observées par mois. Comparez visuellement et quantitativement les deux courbes.
6. Calculez le coefficient de corrélation entre la variable Perruche à collier et chacune des autres variables. Quelle est la plus forte et la plus faible valeur obtenue ? A quelles espèces correspondent ces valeurs extrêmes ? Affichez le nombre d'observations de perruches à collier par mois en fonction du nombre d'observations par mois de l'espèce la fortement et faiblement corrélée. Sur chaque figure, représentez chaque mois par une couleur différente, et ajoutez une légende. Commentez.

c) Analyse en Composante Principales (ACP)

7. Appliquez un ACP. Affichez les deux premières composantes principales sous forme d'un nuage de points pour visualiser les résultats. Identifiez chaque mois par une couleur différente, et ajoutez une légende. Ajoutez aussi sur les axes le pourcentage de variance expliquée par chaque composante. Commentez.
8. Affichez la variance de chaque variable obtenue dans l'espace de l'ACP. Commentez. Quelle est la définition de la variance expliquée ? Quelle est le lien avec les variances que vous venez de calculer ?
9. Affichez et commentez le cercle de corrélation. Permet-il de retrouver les résultats des questions 3 et 6 ?
10. Affichez la figure de la question 7, mais en y superposant cette fois-ci le vecteur associé à la variable perruche à collier affiché à la question 9. En déduire le mois où on observe le plus de perruches à collier.

d) Régression linéaire simple

Nous tentons maintenant d'ajuster un modèle de régression linéaire pour prédire le nombre de perruches à collier en fonction du nombre d'oiseaux d'une autre espèce. Le modèle de régression linéaire s'écrit donc simplement :

$$\textit{Perruches à collier} = \beta 0 + \beta 1 * \textit{oiseau} + \varepsilon \qquad (1)$$

Pour commencer, nous allons effectuer une régression linéaire entre la variable *Perruche à collier* et celle associée à la valeur maximale de corrélation obtenue à la question 6.

11. Appliquez une régression linéaire. Quelle est la valeur du coefficient de détermination R2 – ajusté et non ajusté ? Donnez aussi les valeurs de β0 et β1 prédites par le modèle. Analysez quantitativement et visuellement vos résultats.

12. Calculez l'intervalle de confiance à 90% pour β0 et β1. Interprétez les résultats.

13. Evaluez l'hypothèse de pente nulle pour le coefficient β1 et concluez sur l'existence d'une relation linéaire entre les deux variables. Le coefficient β1 est-il significativement non nul ?

14. Reprenez les questions 11 à 13 en effectuant une régression linéaire entre la variable *Perruche à collier* et celle associée à la valeur minimale de corrélation obtenue à la question 6.

d) Régression linéaire multivariée

Nous allons maintenant ajuster plusieurs modèles de régression linéaire afin de prédire la relation existante entre la cible *Perruche à collier* et les autres variables.

15. Combien y a-t-il de combinaisons possibles de variables – en excluant la variable *Perruche à collier* ? Appliquez une régression linéaire entre la variable cible *Perruche à collier* et chaque combinaison possible de variables autres. Quelle est la valeur maximale que vous obtenez pour le coefficient de détermination ajusté ? Affichez les variables correspondantes, ainsi que les paramètres βi obtenus. Commentez.

16. Déterminez expérimentalement le nombre de modèles pour lesquels on observe une relation linéaire entre la variable cible et les autres variables sélectionnées. Reprenez la question 15 en considérant uniquement ces modèles.

III. Références

- https://www.lpo.fr/la-lpo-en-actions/connaissance-des-especes-sauvages/suivis-ornithologiques
- https://www.lpo.fr/decouvrir-la-nature/fiches-especes/fiches-especes/oiseaux/perruche-a-collier