Analyse des flux de mobilité douce à Bruxelles : classification temporelle et modélisation des passages de la station de comptage CB02411

Géry Bochenski

Contexte

Dans le cadre du plan régional de mobilité de la région de Bruxelles-Capitale, une équipe de chercheurs de l'UCLouvain s'intéresse à l'impact du climat urbain sur l'usage des pistes cyclables à Bruxelles. Pour mener à bien cette étude, une base de données a été mise à disposition par le programme Good Move de la région bruxelloise, regroupant les mesures de fréquentation cycliste de 2022 à 2024 issues de 18 bornes de comptage réparties stratégiquement à travers la capitale.

Ces données sont complétées par plusieurs variables explicatives issues de l'Observatoire Royal de Météorologie et d'Astronomie de Uccle. L'objectif est de voir dans quelle mesure ces facteurs influencent la fréquentation cycliste, et d'évaluer si le programme lancé en 2020 est bien engagé pour atteindre son objectif de quadrupler la mobilité douce d'ici 2030.

Contours du travail

Ce projet se décline en deux grandes parties :

Dans un premier temps, l'analyse s'est portée sur le profil mensuel des comptages vélo par station, sur les jours de la semaine et sur les années 2022 à 2024. Une analyse en composantes principales (ACP) suivie d'un clustering a permis d'identifier plusieurs profils types de stations, en mettant en évidence la particularité de la station CB02411 selon les moments de la journée (jour : 7h–19h / nuit) et les variations mensuelles.

La seconde partie se concentre sur une borne spécifique : CB02411. L'idée est de modéliser son évolution quotidienne du nombre de passages à l'aide d'un modèle de régression multiple en y intégrant des variables comme le jour de la semaine, les indicateurs calendaires et les données météorologiques

Enfin, un tableau de bord interactif sera mis en place pour visualiser et valoriser les résultats clés de l'analyse.

Observations principales

L'analyse multivariée menée sur les stations de comptage met en évidence plusieurs groupes de profils cyclistes, différenciés selon les périodes (jour/nuit) et les jours de la semaine. La station CB02411 se distingue nettement par une fréquentation élevée et régulière, de jour comme de nuit, atteignant jusqu'à 24 000 passages nocturnes en moyenne en juin. Son positionnement excentré dans le plan factoriel et son appartenance aux clusters les plus fréquentés confirment son rôle central dans le réseau cyclable, en lien avec sa localisation stratégique sur le canal de Willebroek à l'entrée de Bruxelles.

Les modèles de régression multiple confirment ce profil, en montrant que la fréquentation est principalement influencée par les variables calendaires (forte hausse en semaine, baisse les week-ends, jours fériés et vacances scolaires) et les conditions météorologiques (effet positif de la température et de l'ensoleillement, effet négatif de la pluie, du vent et de la nébulosité). La fréquentation a par ailleurs nettement progressé entre 2022 et 2024, traduisant une dynamique favorable au vélo urbain. L'ensemble de ces résultats souligne un usage intensif de ce tronçon et montre que la fréquentation est sensible aux variations climatiques et temporelles.

Introduction aux données

Les données ont été téléchargées sur une base de données MySQL (BikeClimate) et intégrées directement dans le logiciel R en utilisant les packages MariaDB et data.table. Les données utilisées sont :

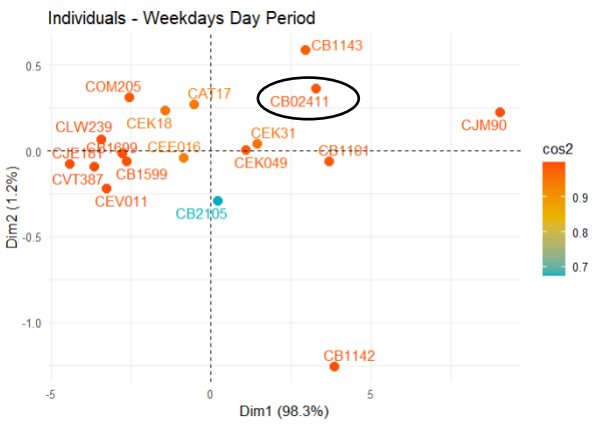
1. Une table de donnée des 18 bornes de comptage cycliste de la région bruxelloise contenant le nom des stations *(FeatureId)*, la date en format UTC1 et CET *(DateUTC1 et DateCET)*, les heures en format UTC1 et CET *(hourUTC1 et hourCET)*, le comptage par heure du nombre de vélo *(Count)*, la vitesse moyenne pondérée par heure *(Speed)*, la température moyenne radiante *(Tmrt) et* l'indice climatique universel de confort thermique *(UTCI)*.
2. Une seconde table de données issue de l'Observatoire Royal de Météorologie et d'Astronomie de Uccle pour chaque heure des années 2022 à 2024 avec pour variables : la température de l'air en °C *(Ta_ucc)*, la vitesse du vent en m/s *(wind_speed_ucc),* la pression atmosphérique en hPa *(pressure_ucc)*, la nébulosité *(cloud_ucc)*, la précipitation en mm lissée toutes les 6 heures *(rain_ucc)* et le degré d'ensoleillement mesurée par la puissance des panneaux solaires *(solar_bxl)*.
3. Un troisième fichier au format .xlsx, intitulé *Vacances FR+NL*, a été importé. Il contient deux variables binaires, Students_Holidays_FR et Students_Holidays_NL, indiquant si les étudiants sont en vacances ou non *(1 = vacances, 0 = hors vacances)*.

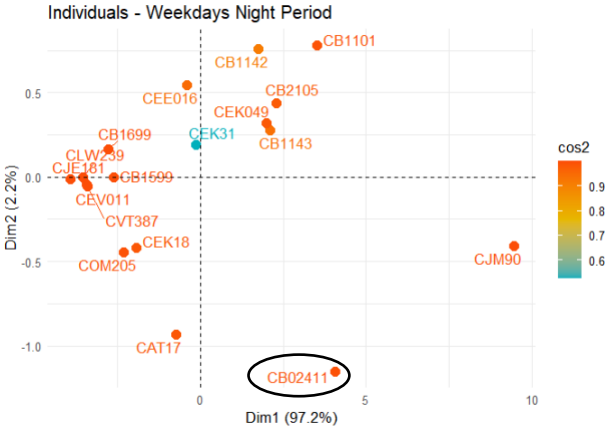
Partie 1 - Analyse en composante principale

Cette section présente les résultats de l'analyse en composantes principales (ACP) appliqués aux données de comptage cycliste de la station CB02411 sur les jours ouvrables. Deux ACP distinctes ont été menées, la première sur les moyennes mensuelles de passages en journée (7h-19h), et la seconde sur les moyennes mensuelles la nuit (19h-7h). Dans les deux ACP (jour et nuit), la première dimension (Dim1) concentre l'essentiel de la variance (98.3 % en journée, 97.2 % la nuit).

Le cercle des corrélations montre que toutes les variables mensuelles sont fortement représentées sur cet axe avec des cos² supérieurs à 0.99.

Le plan factoriel montre que la majorité des stations est concentrée sur la gauche de la Dimension 1, traduisant des profils mensuels de fréquentation relativement similaires. Cependant, la station CB02411 se distingue dans les deux périodes. En journée, elle se positionne à droite sur Dim1, révélant une forte activité. En période nocturne, elle conserve sa position à droite sur Dim1 (activité nocturne soutenue) mais descend fortement sur Dim2. Cela indique un profil spécifique, potentiellement lié à une fréquentation plus marquée à certains moments du mois ou de la semaine.





Clustering

Ensuite, un clustering HCPC a été réalisé à partir de ces ACP. Voici les résultats :

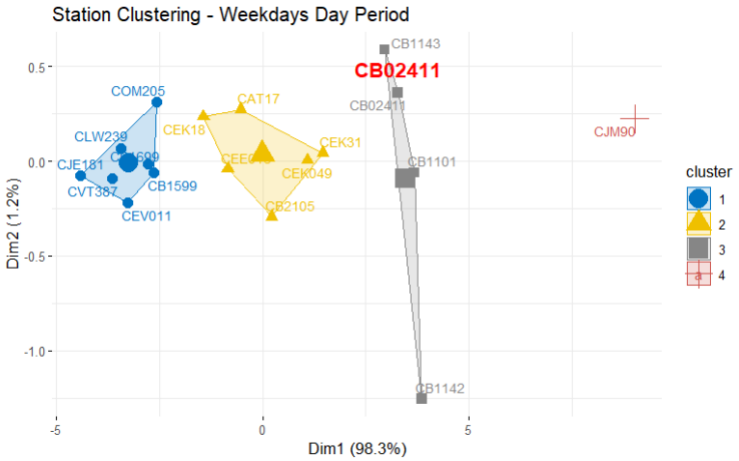
1. *L'analyse pour la période 7h – 19h* a révélé 4 clusters avec des typologies différentes.

Tableau comparatif des passages cyclistes en juin (2022–2024) sur jours ouvrables, de 7h à 19h

Station	Classification	Total passage juin - Jours ouvrables 2022 à 2024 (7h-19h)
COM205	CLUSTER 1	64 871
CAT17	CLUSTER 2	113 259
CB02411	**CLUSTER 3**	**205 429**
CJM90	CLUSTER 4	317 781

Le cluster 1 regroupe les stations à faible fréquentation avec des moyennes mensuelles inférieures à la moyenne globale.

Le Cluster 2 est caractérisé par les stations présentent un profil moyen et sans écart marqué par rapport à la tendance générale. R retourne « NULL » car aucune variable n'a de v.test significatif.



Le cluster 3 comprend les stations à forte fréquentation, dont **CB02411**. Cette dernière se distingue nettement par ses volumes très élevés sur presque tous les mois de l'année.

Le cluster 4 ne contient qu'une seule station *(CJM90)* qui est isolée du reste à cause de volumes extrêmement élevés.

J'ai ensuite dressé un tableau comparatif des stations du cluster 3 afin de comprendre l'origine de cette forte dispersion. **CB02411** présente un volume particulièrement élevé en juin *(205 429 passages)*, avec un profil très marqué entre les saisons. Des stations comme CB1143 *(190 127)* ou CB1142 *(223 893)* ont elles aussi des volumes importants mais leurs variations mensuelles sont différentes, ce qui justifie leur éloignement spatial dans le cluster.

Total du passage mensuel de 2022 à 2024 (Cluster3) pour la période jour et sur les jours ouvrables

Station	Janvier	Avril	Juin	Juillet	Novembre
CB02411	**24683**	**45236**	**72347**	**54343**	**38672**
CB1101	31428	37915	55024	45353	42147
CB1143	24542	26125	48365	38607	38862
CEK049	23949	32095	47728	34954	34290

2. *L'analyse pour la période 19h-7h* a révélé 3 clusters et non 4 comme en journée.

Le cluster 1 regroupe des stations à faible trafic toute l'année.

Le cluster 2 correspond aux stations à forte fréquentation, tandis que le cluster 3 reste composé uniquement de la station CJM90, caractérisée par des volumes exceptionnellement élevés.

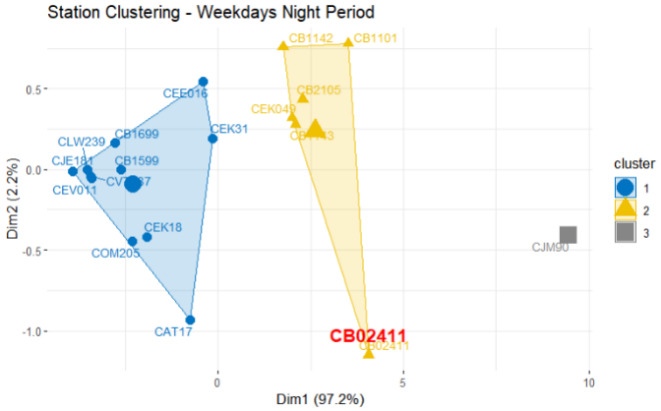


Tableau comparatif des passages cyclistes en juin (2022–2024) sur jours ouvrables, de 19h à 7h

Station	Classification	Total passage juin - Jours ouvrables 2022 à 2024 (19h-7h)
COM205	CLUSTER 1	20 802
CB02411	CLUSTER 2	72 347
CJM90	CLUSTER 3	101 888

La station CB02411 est affectée au cluster 2 mais se situe très à l'écart du barycentre du groupe, à la droite et tout en bas du plan factoriel. Elle est caractérisée par une fréquentation nocturne particulièrement élevée, surtout durant le printemps et l'été. Le tableau ci-dessous, qui présente les moyennes mensuelles de 2022 à 2024 pour quatre stations du cluster 2, illustre cet écart.

Station	Janvier	Avril	Juin	Juillet	Novembre
CB02411	**95391**	**135914**	**205429**	**149425**	**156476**
CB1101	112584	151617	202911	160081	149424
CB1143	98141	119886	190127	148806	152247
CB1142	121870	153161	223893	123621	177526

En conclusion, la station CB02411 apparaît comme un point de comptage à fort trafic tant de jour que de nuit, avec une dynamique saisonnière marquée en particulier lors des mois printaniers et estivaux.

Partie 2 - Nettoyage du dataset CB02411 et imputation des données météo

L'analyse exploratoire a révélé deux cas de valeurs manquantes dans le jeu données CB02411 :

Vitesse (Speed) : 430 valeurs manquantes, principalement liées à des moments où le compteur affichait zéro passage *(Count = 0)*. Ce comportement est cohérent, car aucune vitesse n'est mesurée sans cyclistes.

Comptage (Count) : 16 valeurs manquantes.

- **3 valeurs manquantes** à 23h. Elles sont dues à un décalage horaire non pris en compte (UTC+1, passage hiver/été). Une imputation manuelle par la médiane a été appliquée spécifiquement à ces cas.
- **13 jours** où des valeurs nulles ont été observées durant les périodes du 4–13 mars 2022 et du 18–20 août 2023 coïncident avec des travaux majeurs (KANAL[1], Viaduc de Vilvorde[2])

[1] https://kanal.brussels/sites/default/files/about/kan-rapport_activites_2023_fr_1.pdf
[2] https://sau.brussels/sites/canal

susceptibles d'avoir perturbé localement les flux cyclistes. Ces observations ont été exclues des modèles prédictifs pour éviter un biais contextuel.

2. Imputation (weather_uccle)

Les données météorologiques contenaient plusieurs valeurs manquantes, notamment dans la variable cloud_ucc. Elles apparaissaient souvent par blocs journaliers ce qui me laisse présager un schéma MAR *(Missing At Random)* suggérant une origine structurelle telle que des pannes ou une absence de relevé.

Pour y remédier, une imputation multiple par MICE (*Multivariate Imputation by Chained Equations*) a été mise en œuvre :

- 5 jeux de données complets ont été générés (10 itérations chacun).
- Une régression multiple a ensuite été appliquée sur chaque jeu imputé.
- Les résultats ont été fusionnés avec *pool()* selon les règles de Rubin, intégrant les incertitudes intra et inter-modèles.

Ces étapes garantissent une meilleure robustesse des modèles tout en limitant les biais liés à la perte d'information.

Partie 3 - Modélisation

L'objectif de cette section est de modéliser le nombre total de passages cyclistes journaliers en fonction de variables météorologiques et temporelles, tout en s'appuyant sur les jeux de données issus d'une imputation multiple. Chaque jeu imputé a été fusionné aux données de comptage et agrégé au niveau journalier. Le dataset a ensuite été enrichi avec des variables temporelles telles que les jours fériés et les vacances scolaires en vue d'une modélisation plus complète.

Trois modèles de régression linéaire multiple ont été élaborés afin d'expliquer les variations journalières du trafic cycliste observé à la borne CB02411. Chaque modèle a été ajusté sur les cinq jeux de données issus de l'imputation multiple et définis comme suit :

- **Modèle 1** : Toutes les variables des deux datasets.
- **Modèle 2** : Suppression des variables redondantes et non significatives *(UTCI, press_ucc et cloud_ucc)* et test d'interaction entre « Tmrt » et « season ».
- **Modèle 3** : version simplifiée avec « Tmrt » et sans interaction.Les performances des modèles ont été évaluées à l'aide du R^2 ajusté et du facteur d'inflation de la variance (VIF). Bien que les Modèles 1 et 2 aient obtenu un R^2 légèrement supérieur, ils présentaient une multicolinéarité problématique et trop de variables non significatives. Par conséquent, c'est le modèle 3 sans outliers qui a été retenu pour la suite de l'analyse car il offre un bon équilibre entre performance et robustesse statistique.

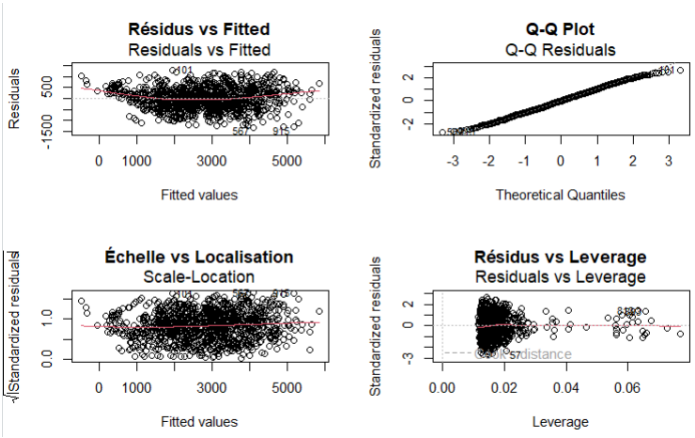
Modèle	Variables	R^2 ajusté	Problème
Model_1	22 variables	0,7943	Multicolinéarité 16/22 significatives Hypothèse de normalité
Model_2	22 variables	0,8067	Multicolinéarité 18/22 significatives Hypothèse de normalité
Model_3	18 variables	0,7937	Hypothèse de normalité 17/18 significatives
Models_3_clean	18 variables	0,8461	

Les valeurs aberrantes persistantes ont été identifiées à l'aide de la distance de Cook (seuil = 4/n) sur les cinq versions imputées. Les observations considérées comme outliers dans au moins trois modèles sur cinq ont été supprimées. Ainsi, 47 outliers ont été retirés du Modèle 3.

Les diagnostics du modèle 3 (sans outliers) montrent que les hypothèses de la régression linéaire sont globalement respectées.

La relation entre les variables explicatives et la variable cible est linéaire, les résidus sont approximativement normaux, bien que de légères déviations existent dans les queues de distribution.

L'homoscédasticité est respectée, sans tendance claire de variation des résidus.



Interpretation des coefficients du modèle

L'intercept du modèle correspond à un lundi d'automne 2022 avec un passage journalier estimé à environ 2772 usagers.

L'analyse montre que le trafic est plus élevé durant la semaine et diminue significativement les week-ends, les jours fériés et pendant les vacances scolaires.

```
> print(sig_table_3)
                    term    estimate  std.error       p.value significance
1            (Intercept)  2772.56120  86.816077 5.438862e-155          ***
2    weekday_nameTuesday   325.72771  56.061928  8.348272e-09          ***
3  weekday_nameWednesday    12.23750  55.445784  8.253617e-01
4   weekday_nameThursday   204.93497  55.887676  2.582261e-04          ***
5     weekday_nameFriday  -357.08894  55.793094  2.365589e-10          ***
6   weekday_nameSaturday -1485.74285  56.038086 3.879300e-118          ***
7     weekday_nameSunday -1398.98752  56.311149 7.747403e-107          ***
8         is_holidayTRUE  -960.45477 118.796349  1.759734e-15          ***
9       student_holiday  -643.70173  36.056155  3.300030e-62          ***
10                Tmrt    83.44504   3.226055 1.090398e-113          ***
11           rain_ucc -1116.66939  79.344286  3.198945e-41          ***
12       wind_speed_ucc   -70.39604  11.974963  5.607437e-09          ***
13           year2023   768.41726  36.901379  1.833792e-80          ***
14           year2024  1137.71781  37.385168 4.285782e-145          ***
15       seasonspring  -516.18272  48.717921  5.964834e-25          ***
16       seasonsummer  -661.95071  63.819955  5.008805e-24          ***
17       seasonwinter  -652.15014  44.789677  9.330717e-44          ***
18          cloud_ucc   -45.96364   8.686295  1.500567e-07          ***
```

Les conditions météorologiques influencent fortement la fréquentation : la température et l'ensoleillement favorisent les déplacements, tandis que la pluie, le vent et la couverture nuageuse les réduisent. Enfin, une tendance annuelle à la hausse du trafic cycliste est clairement observée entre 2022 et 2024, ce qui pourrait refléter les effets positifs des politiques publiques mises en place en faveur de la mobilité durable et de l'aménagement cyclable à Bruxelles.

Conclusion

L'objectif de ce travail était de modéliser les passages cyclistes journaliers à la borne CB02411 à Bruxelles en fonction de facteurs temporels et météorologiques. Le modèle final, ajusté après imputation multiple et exclusion des outliers explique environ 84 % de la variance du trafic cycliste journalier à la station CB02411. Les analyses mutlivariées ont permis de classer cette station dans un groupe à forte fréquention de jour comme de nuit, particulièrement durant les périodes printanières et estivales. Le modèle final a été évalué au moyen d'indicateur de performance comme l'écart-type résiduel qui mesure la différence entre les valeurs réelles et les valeurs prédites de mon modèle. L'erreur est de 479,03 passages par jour (soit 16,94 %), ce qui est encore raisonnable compte tenu de la variabilité du traffic et de sa dépendance aux facteurs extérieurs (météo, événements,etc..).