Anticiper les besoins en consommation électrique des bâtiments

Soutenance Projet 4
Parcours Data Scientist
Ilaria Mereu

Problématique

Seattle

Prédire les émissions de gaz à effet de serre et la consommation totale d'énergie des bâtiments non destinés à l'habitation

Source des données: les données déclaratives des propriétés immobilières de la ville de Seattle comprenant leurs caractéristiques et leur consommation énergétique en 2015 et 2016.

Nos objectifs:

1) Tenter de prédire les émissions de gaz à effet de serre et la consommation totale d'énergie des bâtiments non destinés à l'habitation.

2) Évaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions, qui est fastidieux à calculer avec l'approche utilisée actuellement.

Sommaire

1. Description des jeux de données d'origine
2. Analyse exploratoire, feature engineering et stratégie
3. Modèles explorés et résultats
4. Intérêt de l'Energy Star Score
5. Perspectives

1 - Description des jeux de données d'origine

Description des deux jeux de données

- 2 jeux de données se référant aux années 2015 et 2016
  - 2015: 3340 lignes x 47 colonnes, dont aucune vide
  - 2016: 3376 lignes x 46 colonnes, dont une colonne vide → 3376 lignes x 45 colonnes non vides
- À chaque ligne correspond un regroupement immobilier identifié avec un code cadastral

- Les informations sur l'unité immobilière: code cadastral, nom de l'exercice, **nombre de bâtiments, nombre d'étages, année de construction, surfaces couvertes et espaces de parking**, utilisation de la structure.
- Les détails de la **consommation énergétique et des émissions de CO_2 équivalente** des immeubles.

Variables target

Consommation énergétique: ‘SiteEnergyUse(kBtu)’ et son logarithme

La quantité annuelle d’énergie consommée par la propriété à partir de tout type de sources énergétiques. Elle est mesurée milliers d’unités thermales britanniques (kBtu).

Émissions de gaz à effet de serre : ‘TotalGHGEmissions' et son logarithme

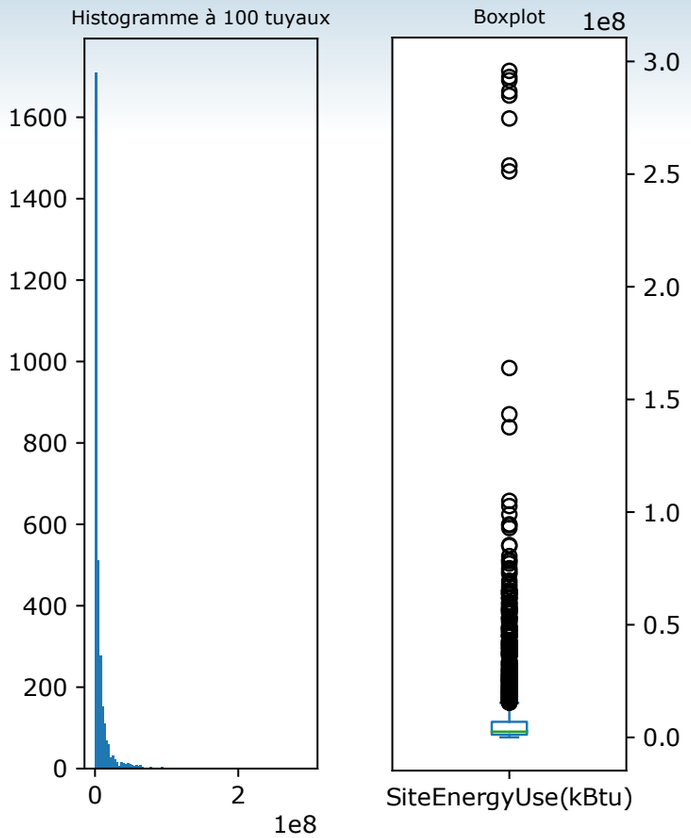
La quantité totale d’émissions de gaz à effet de serre (CO_2, méthane, etc.) diffusée dans l’atmosphère à cause de la consommation énergétique de la propriété. Elle est mesurée en tonnes de CO_2 équivalentes.

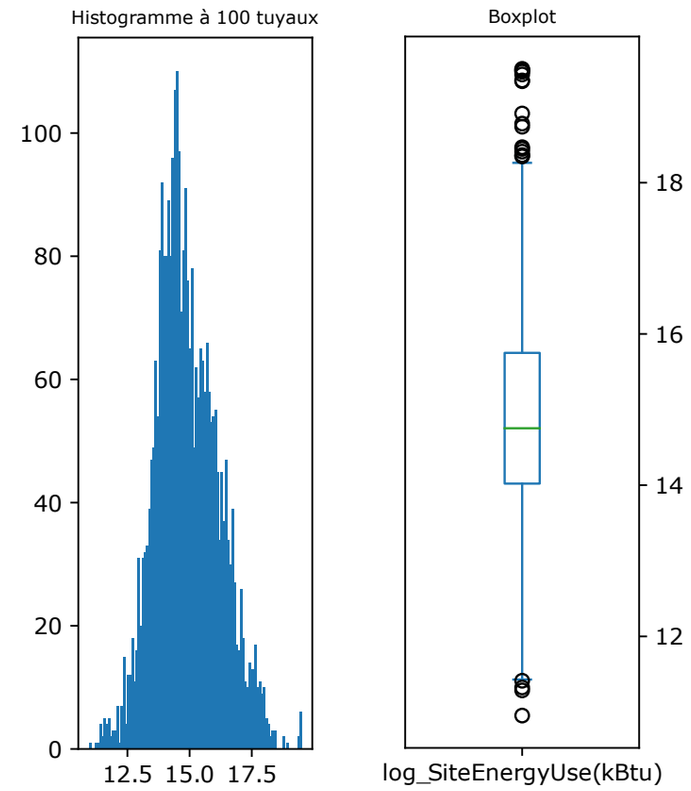
https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking?select=socrata_metadata_2015-building-energy-benchmarking.json

Variables target

Consommation énergétique: ‘SiteEnergyUse(kBtu)’ et son logarithme

La quantité annuelle d’énergie consommée par la propriété à partir de tout type de sources énergétiques. Elle est mesurée milliers d’unités thermales britanniques (kBtu).

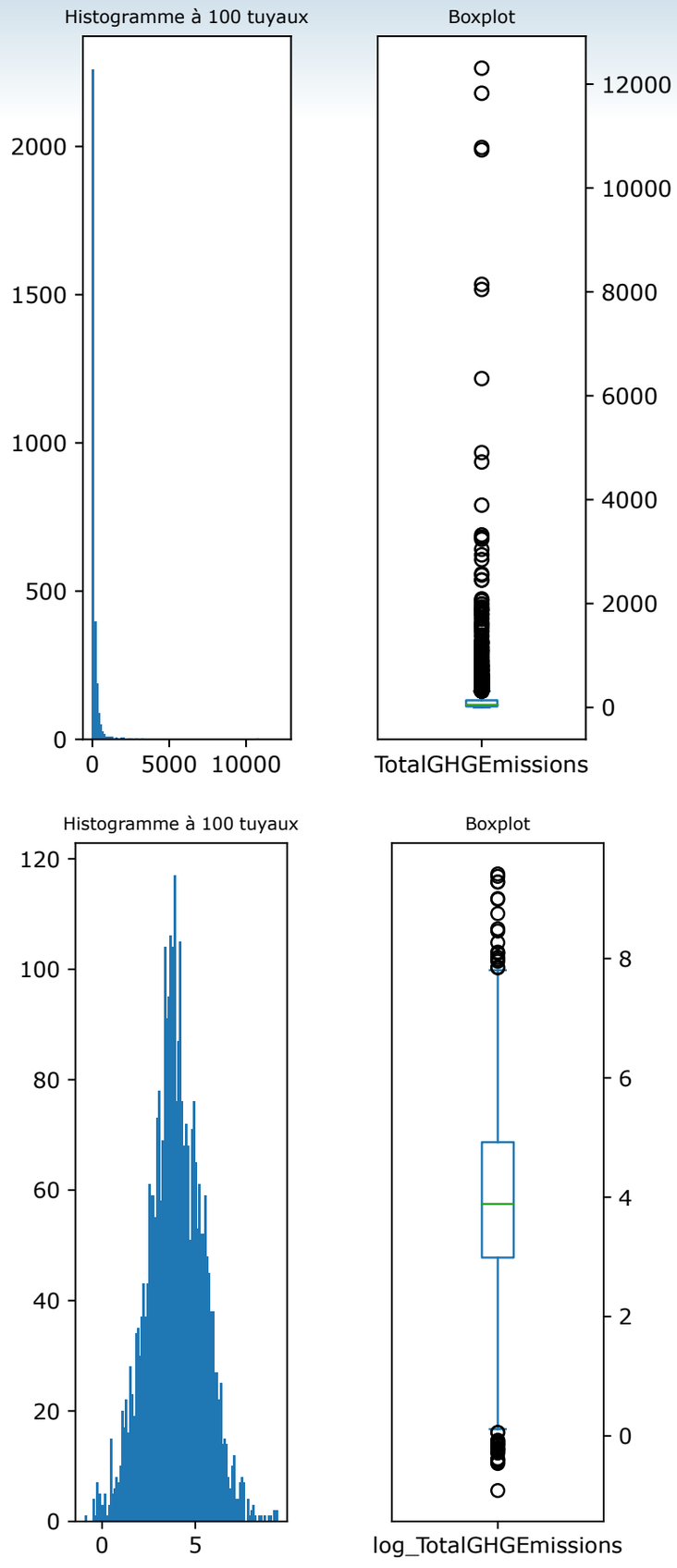




Variables target

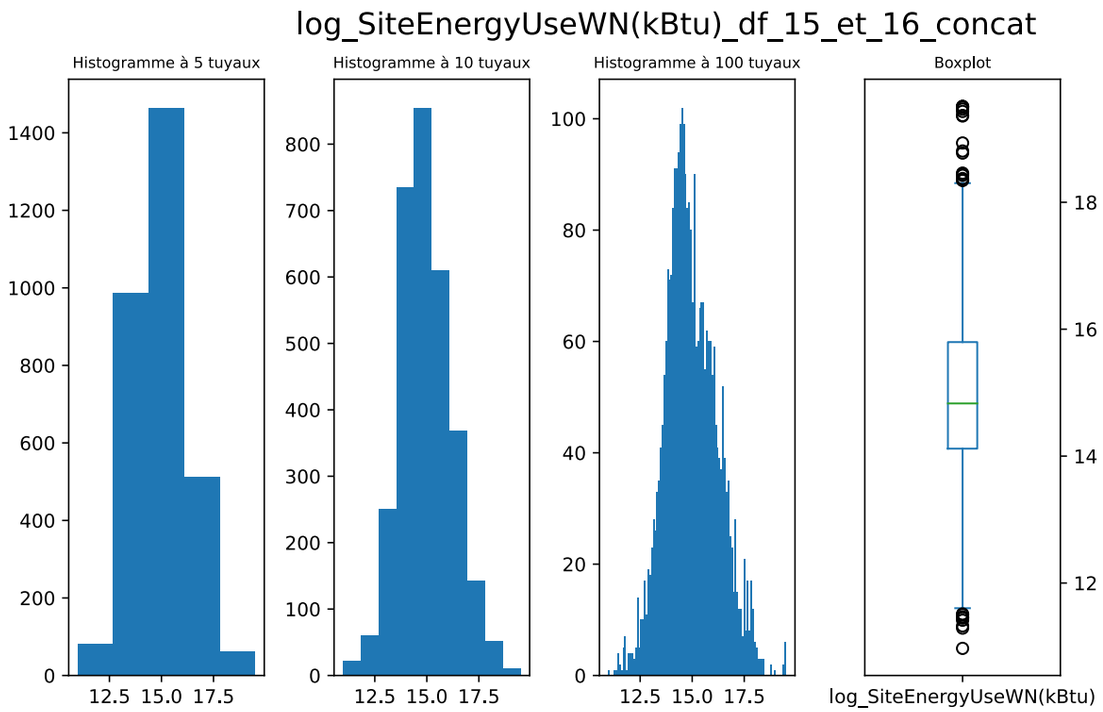
Émissions en gaz de serre :
'TotalGHGEmissions' et son logarithme

La quantité totale d'émissions de gaz de serre (CO_2, méthane, etc.) diffusée dans l'atmosphère à cause de la consommation énergétique de la propriété. Elle est mesurée en tonnes de CO_2 équivalentes.



Études de normalité

Consommation énergétique: logarithme de ‘SiteEnergyUse(kBtu)’



Sur le jeu de données nettoyé:

3108 values

Mode(s):[10.97016530575204, 11.289380563625867, 11.324473153445949, 11.413753572773516, '...']
Moyenne = 14.9628
Écart type = 1.2628
Mediane = 14.8300

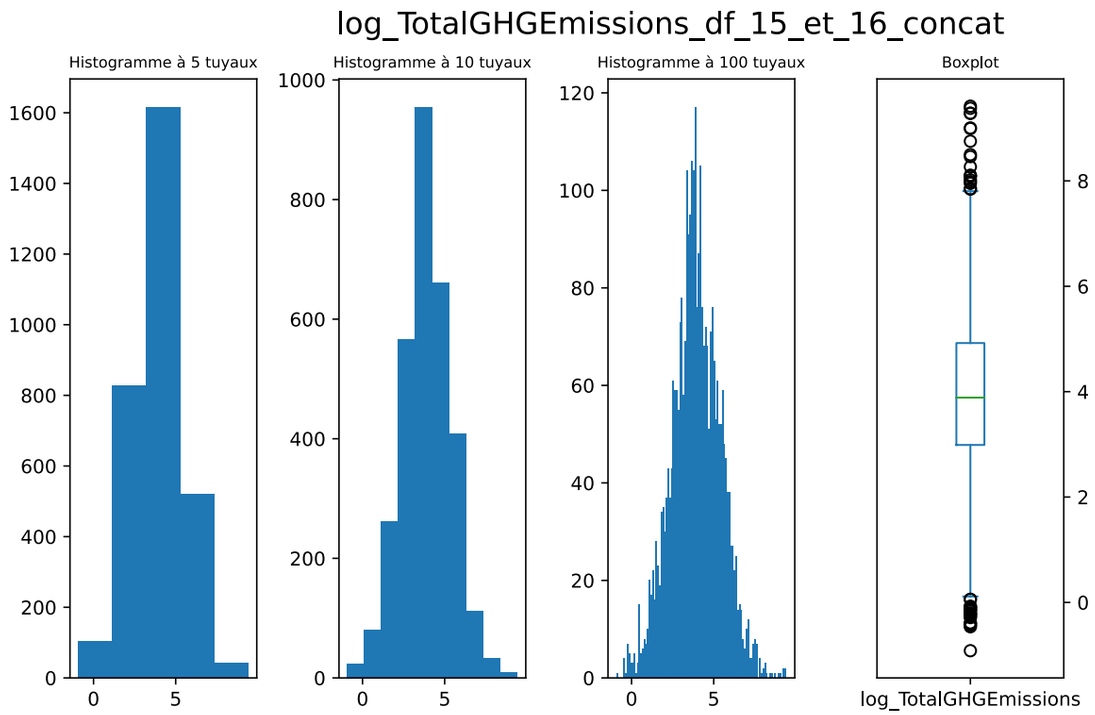
Test de normalité de Anderson

Statistique=8.60
L'hypothèse nulle que les données proviennent de la distribution normale
peut être rejetée avec un niveau de signifiance de 2.5%.

Valeurs critiques=[0.575 0.655 0.786 0.917]
Niveau de signifiance=[15. 10. 5. 2.5]]

Études de normalité

Émissions de gaz de serre : logarithme de ‘TotalGHGEmissions' et son logarithme



Sur le jeu de données nettoyé:

3108 values

Mode(s):[1.8180767775454283, 1.8229350866965048, 1.840549633397487, 1.9035989509835904, '...']
Moyenne = 3.9250
Écart type = 1.4852
Mediane = 3.8867

Test de normalité de Anderson

Statistique=1.51
L'hypothèse nulle que les données proviennent de la distribution normale
peut être rejetée avec un niveau de signifiance de 2.5%.

Valeurs critiques=[0.575 0.655 0.786 0.917]
Niveau de signifiance=[15. 10.   5.   2.5]]

2 - Analyse exploratoire et stratégie

Opérations sur le jeu de données

- Certaines colonnes ont été renommées afin de donner plus de conformité aux deux jeux de données:
  - 2015:
    - 'GHGEmissions(MetricTonsCO2e)' → 'TotalGHGEmissions'
    - 'GHGEmissionsIntensity(kgCO2e/ft2)' →'GHGEmissionsIntensity'
    - 'Seattle Police Department Micro Community Policing Plan Areas' → 'SPD MCPP Areas'

- Suppression des lignes et des colonnes vides ou aberrantes*
  - 2015: 3340 lignes x 47 colonnes → 3310 lignes x 47 colonnes
  - 2016: 3376 lignes x 46 colonnes → 3331 lignes x 45 colonnes

  * (valeurs égales ou inférieures à zero pour le colonnes relatives à des surfaces ou aux consommations: 'PropertyGFABuilding(s)', 'SiteEUI(kBtu/sf)', 'SiteEUIWN(kBtu/sf)', 'TotalGHGEmissions', 'GHGEmissionsIntensity')

- Sélection des bâtiments non résidentiels (ajout de la colonne "NonRes")
  - 2015: 3340 → 1602 lignes x 48 colonnes
  - 2016: 3376 → 1613 lignes x 46 colonnes

- Conformité aux standard énergétiques: ajout de la colonne "Compl" sans sélection
  - 2015: → 1602 propriétés résultent conformes
  - 2016: → 1513 propriétés résultent conformes

- On renomme la notation des outliers pour la rendre homogène: ajout de la colonne ['Outlier_ok']

- Ancienneté des propriétés : ajout de la "Anciennete". Dans cette colonne, on note l'âge de la propriété par rapport à l'année de référence du jeu de données (2015 ou 2016). L'année de construction est repérable dans la colonne 'YearBuilt'.

- Estimation de la surface extérieure : ajout de la colonne "measure_floors" (voir à la suite)

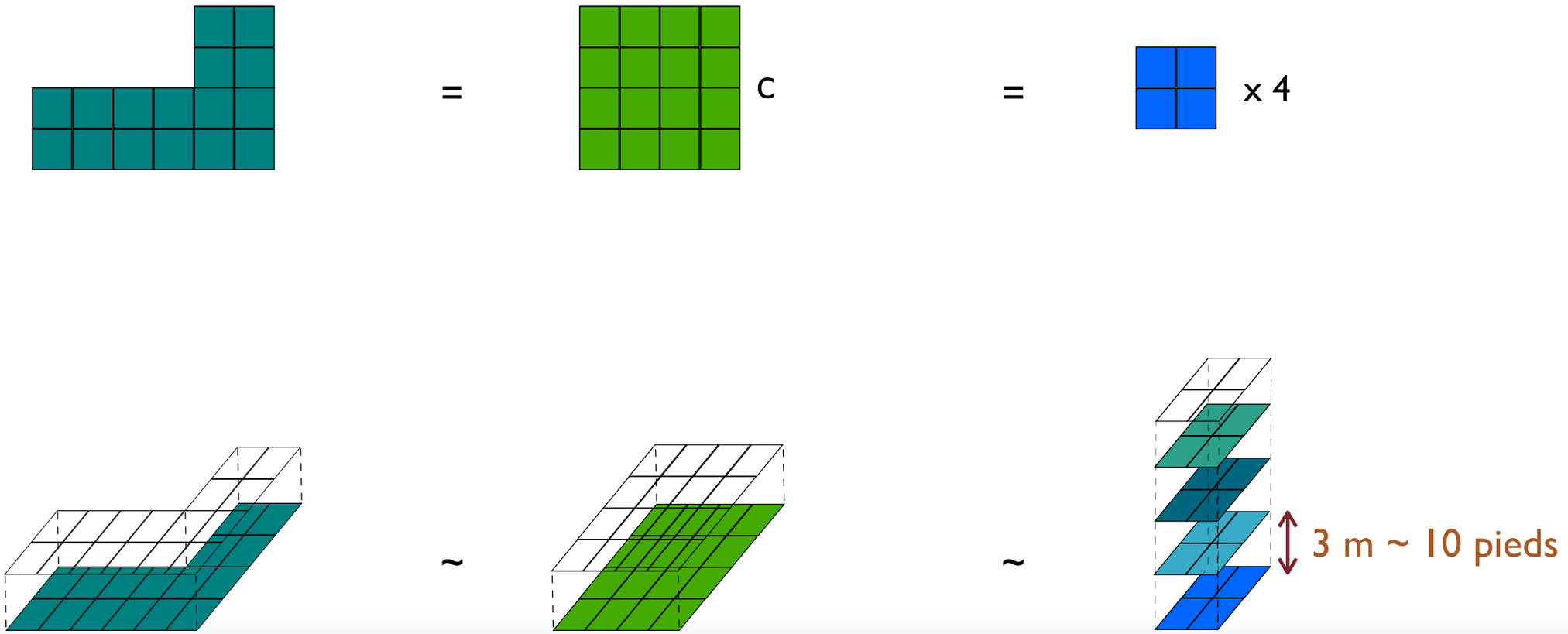
- Normalisation

2015 → 1602 lignes x 52 colonnes
2016 → 1613 lignes x 50 colonnes

Estimation de la surface extérieure

La colonne “measure_floors”, contient une estimation de la surface extérieure d’une propriété. Plus cette surface est étendue, plus grand sera l’échange avec l’extérieur et donc la consommation énergétique associée à la climatisation de l’intérieur, à parité de toute autre caractéristique.

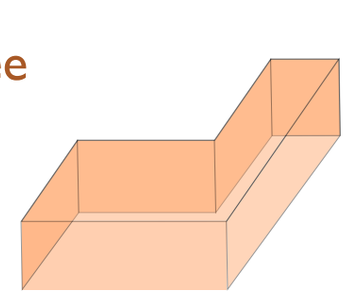
Surface couverte (PropertyGFABuilding(s))

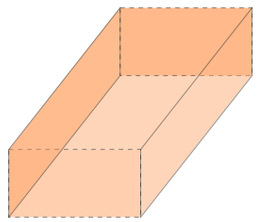


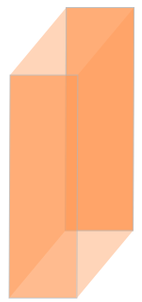
Estimation de la surface extérieure

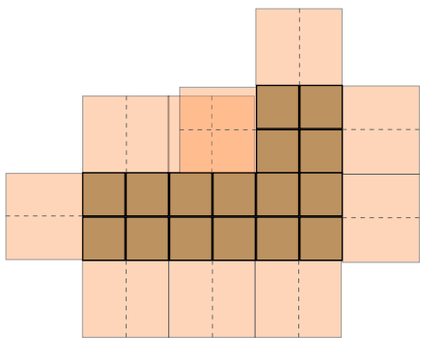
```
df['measure_floors'] = df['PropertyGFABuilding(s)'] / df['NumberofFloors'] +
np.sqrt(df['PropertyGFABuilding(s)'] / df[‘NumberofFloors’]) * 4 * 10 * (df[‘NumberofFloors’])
```

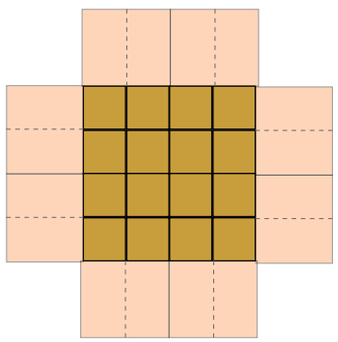
Surface exposée

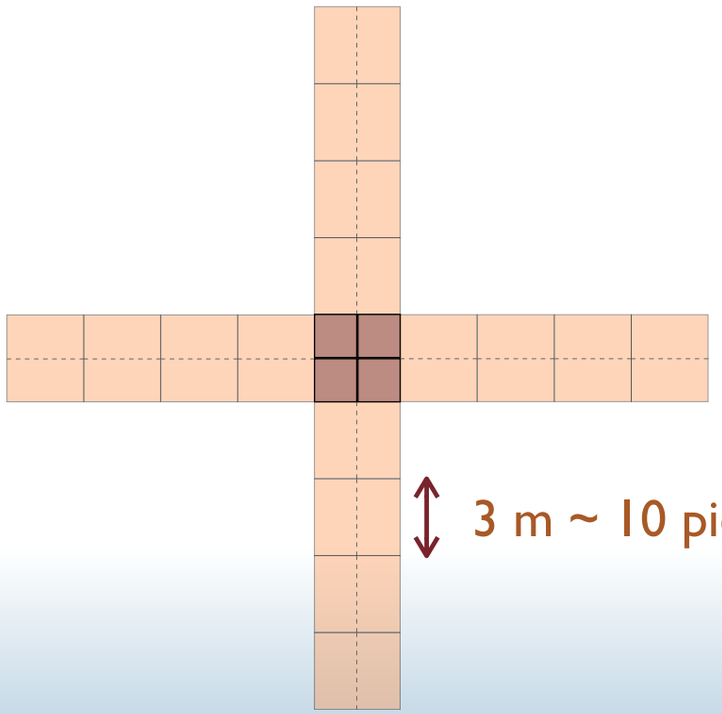






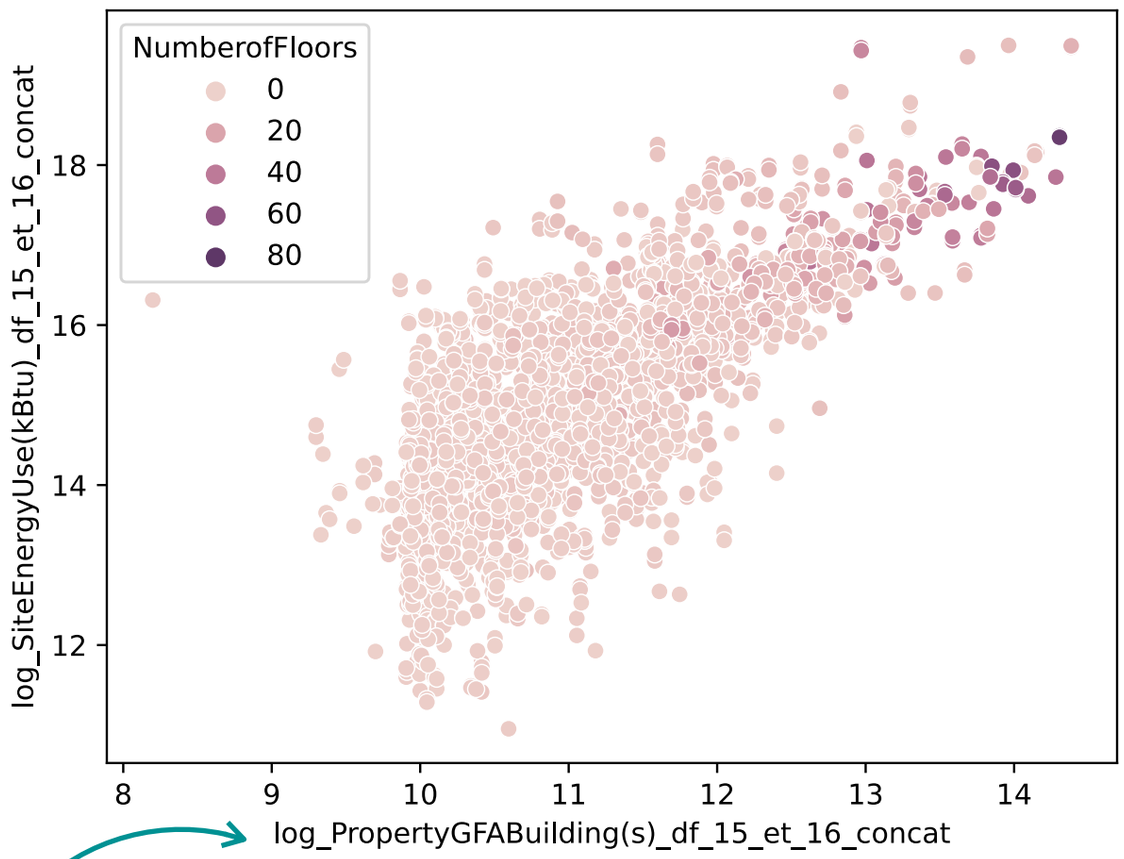






Analyses exploratoires bivariées

log_PropertyGFABuilding(s)_df_15_et_16_concat VS log_SiteEnergyUse(kBtu)_df_15_et_16_concat - hue_NumberofFloors

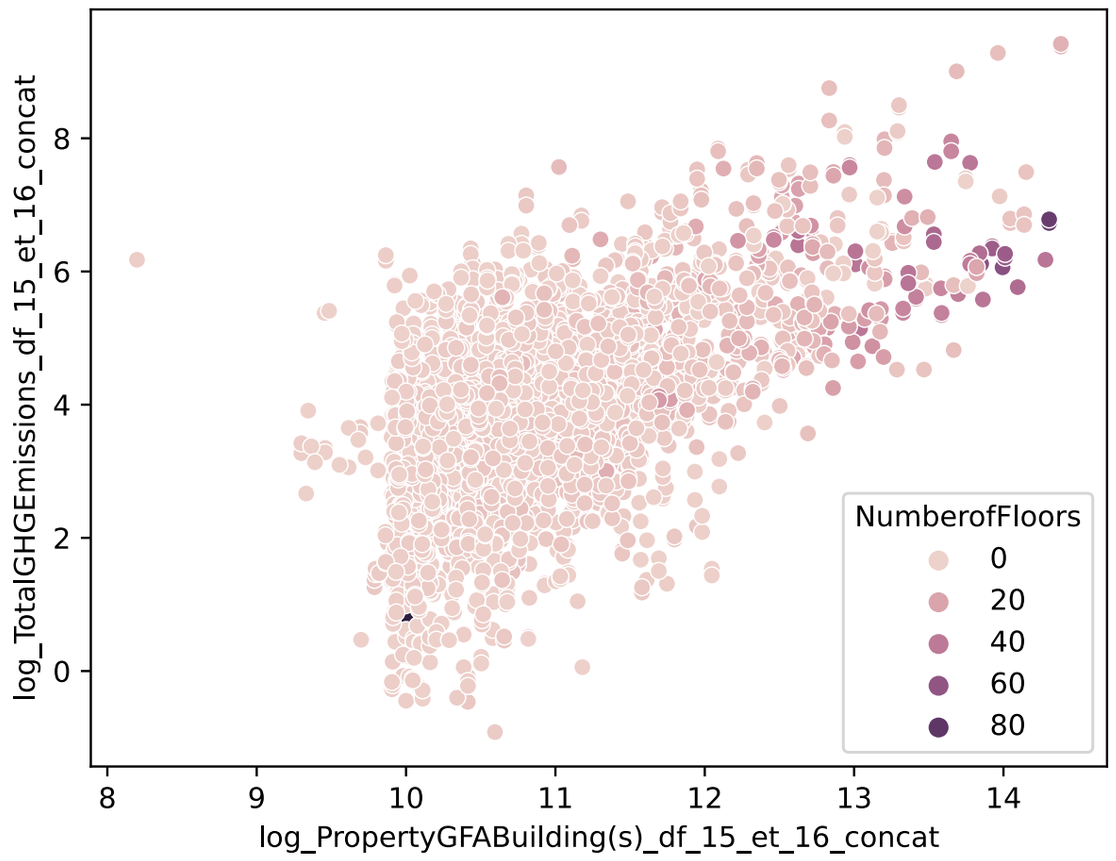


Coefficient de Pearson: 0.75; p-value = 0.00

Les analyses bivariées des logarithmes de variables extensives telles que la consommation totale (log SiteEnergyUse(kBtu)) et la surface couverte (log PropertyGFABuildings) suggèrent une relation d'ordre linéaire entre ces type de variables.

Analyses exploratoires bivariées

log_PropertyGFABuilding(s)_df_15_et_16_concat VS log_TotalGHGEmissions_df_15_et_16_concat - hue_NumberofFloors



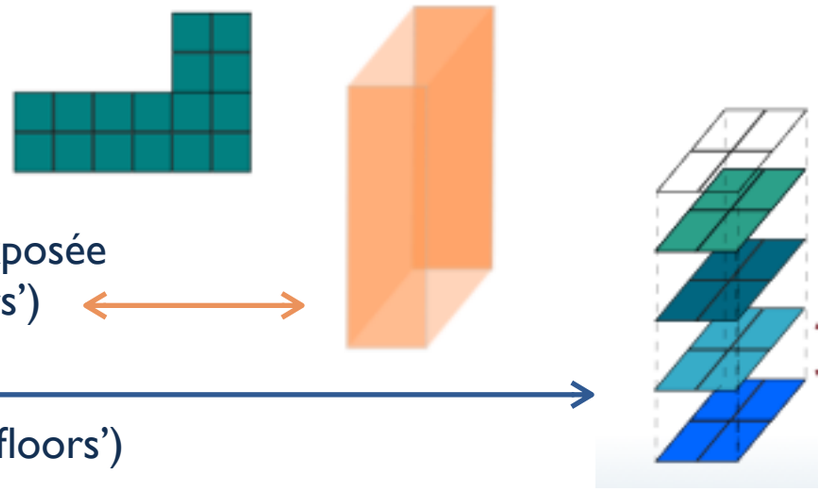
Coefficient de Pearson: 0.61; p-value = 0.00

Dans le cas des émissions, les indices de linéarité sont moins prononcés : un trait récurrent dans cette analyse.

Features inclues dans les modèles

Pour définir les possibles modèles, nous avons sélectionné 5 ou 3 features parmi:

- les mesures de surface ('PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuildings')
- L'estimation de la surface exposée à l'extérieur ('measure floors')
- Le nombre d'étages (indirectement en 'measure floors')
- Le nombre de bâtiments ('NumberofBuildings') composant la propriété
- l'âge des immeubles ('Anciennete')



Stratégie finalisée à limiter l'overfitting

Points critiques de l'approche simple train_split + gridsearch

1 x train_split

Haute dépendance du jeu sélectionné pour l'entrainement du modèle

→

1 x Gridsearch

Forte variabilité des hyperparamètres sélectionnés d'un run à l'autre

→

Overfitting / manque de généralité

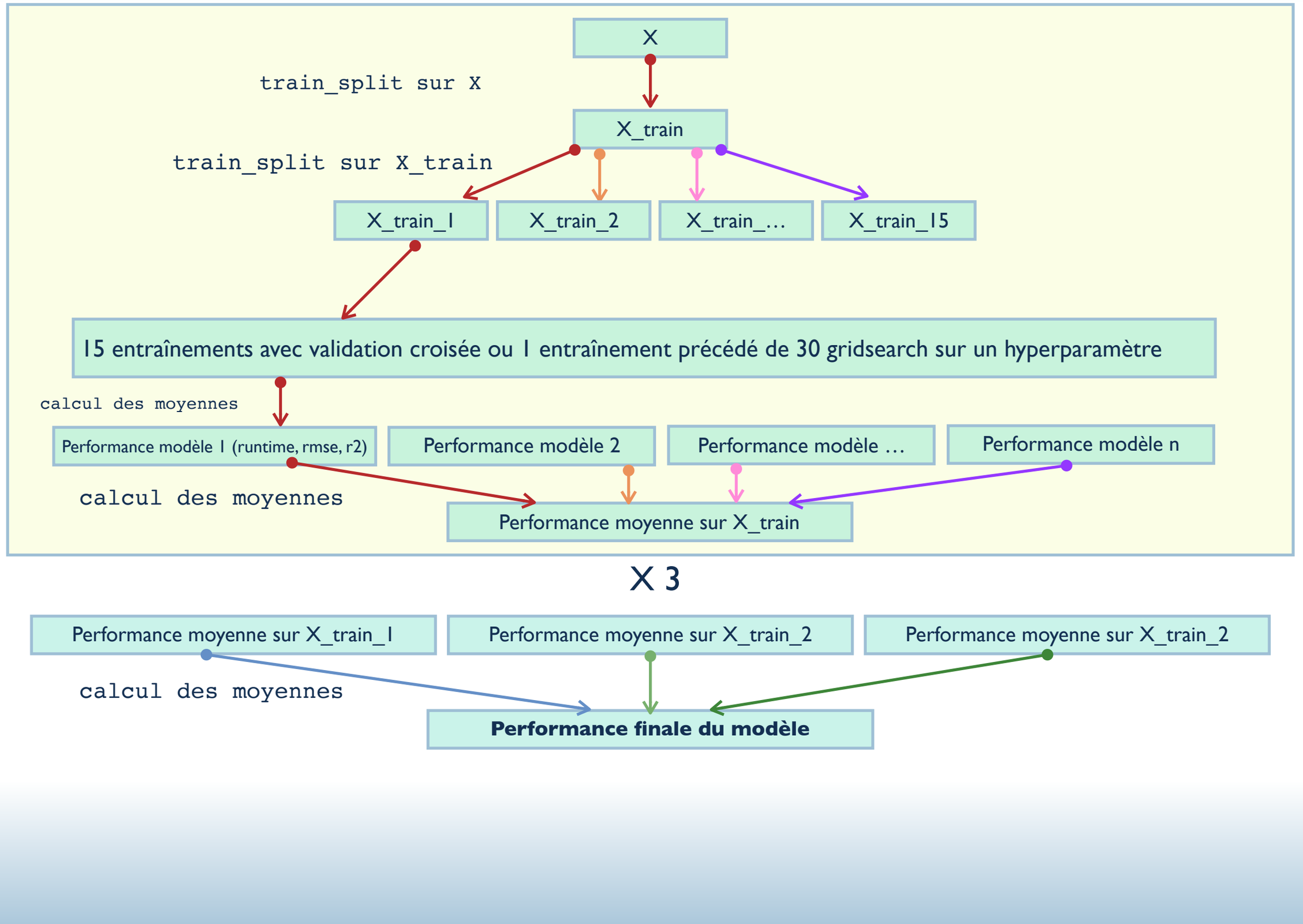
Approche choisie

- Entraînements multiples sur différentes sections de X
- Moyennes

- Gridsearch multiples
- Moyennes

```
END best_alpha_rmse_1 and best_alpha_r2_1 (0, 'Ridge ', 3.1622776601683795e-05, 3.1622776601683795e-05)
END best_alpha_rmse_1 and best_alpha_r2_1 (1, 'Ridge ', 100.0, 100.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (2, 'Ridge ', 316.22776601683796, 316.22776601683796)
END best_alpha_rmse_1 and best_alpha_r2_1 (3, 'Ridge ', 100.0, 100.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (4, 'Ridge ', 316.22776601683796, 316.22776601683796)
END best_alpha_rmse_1 and best_alpha_r2_1 (0, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (1, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (2, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (3, 'Lasso ', 31.622776601683793, 31.622776601683793)
END best_alpha_rmse_1 and best_alpha_r2_1 (4, 'Lasso ', 31.622776601683793, 31.622776601683793)
```

Stratégie finalisée à limiter l'overfitting



3. Modèles explorés et résultats

Consommation, variables testées

- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'log PropertyGFABuilding(s)', 'NumberofBuildings', 'Anciennete']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFABuilding(s)', 'Anciennete']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']
- SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFABuilding(s)', 'Anciennete']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'log PropertyGFABuilding(s)', 'NumberofBuildings', 'Anciennete']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'log PropertyGFABuilding(s)', 'Anciennete']
- log SiteEnergyUse(kBtu) VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']

Estimateurs testés:

- Dummy
- Linear
- Ridge
- Lasso
- Elastic Net
- Random Forest

Consommation, 5 features

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	12756243.98	-0.02	15284513.78	-0.00
Linear	0.00	9272277.22	0.48	10791814.32	0.50
Ridge	0.38	13142430.30	0.43	9918432.84	0.50
Lasso	151.84	9527642.74	0.53	10714142.05	0.49
Elastic Net	2.60	12175987.00	0.40	9987757.42	0.53
RF	8.87	6191349.82	**0.82**	4090672.46	**0.92**

Target: SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, log PropertyGFAParking, log PropertyGFABuildings, NumberofBuildings, Anciennete]

- Le passage au logarithme de la variable target est utile pour les estimateurs: linéaire, Ridge et Lasso.
- L'estimateur Random Forest n'en est pas influencé significativement et il est en tout cas le plus performant.

- r2 généralement élevés ≳ 0.4

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.15	-0.00	1.13	-0.00
Linear	0.00	0.69	**0.67**	0.64	**0.67**
Ridge	0.42	0.60	**0.71**	0.66	**0.66**
Lasso	0.48	0.66	0.62	0.66	0.68
Elastic Net	0.07	0.61	**0.70**	0.66	**0.66**
RF	9.08	0.52	**0.80**	0.34	**0.91**

Target: log SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, log PropertyGFAParking, log PropertyGFABuildings, NumberofBuildings, Anciennete]

Consommation, 3 features

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	14149557.84	-0.00	14960188.94	-0.00
Linear	0.00	11087179.48	0.46	10124509.85	0.53
Ridge	0.35	11693647.85	0.35	9970659.39	0.55
Lasso	70.67	8540297.95	0.59	10713650.74	0.50
Elastic Net	21.34	11615400.09	0.46	9946012.39	0.53
RF	7.42	6634264.22	**0.83**	4559092.22	**0.90**

Target: SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, Anciennete, log measure floors]

- Le passage au logarithme de la variable target est utile pour les estimateurs : linéaire, Ridge, Lasso et Elastic Net.
- L'estimateur Random Forest n'en est pas influencé significativement et il est en tout cas le plus performant.

- r2 généralement élevés ≳ 0.4

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.21	-0.00	1.11	-0.00
Linear	0.00	0.69	**0.65**	0.64	**0.68**
Ridge	0.35	0.63	**0.69**	0.66	**0.67**
Lasso	0.36	0.56	**0.71**	0.67	**0.66**
Elastic Net	0.07	0.59	**0.71**	0.67	**0.66**
RF	7.33	0.48	**0.81**	0.30	**0.93**

Target: log SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, Anciennete, log measure floors]

Modèle retenu pour la prédiction de la consommation

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.21	-0.00	1.11	-0.00
Linear	0.00	0.69	**0.65**	0.64	**0.68**
Ridge	0.35	0.63	**0.69**	0.66	**0.67**
Lasso	0.36	0.56	**0.71**	0.67	**0.66**
Elastic Net	0.07	0.59	**0.71**	0.67	**0.66**
RF	7.33	0.48	**0.81**	0.30	**0.93**

Pouvoir prédictif > 80%

- Target: log SiteEnergyUse(kBtu)
- Features: [log PropertyGFATotal, Anciennete, log measure floors]

Émissions, variables testées

- TotalGHGEmissions VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- TotalGHGEmissions VS ['PropertyGFATotal', 'PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'measure floors']
- TotalGHGEmissions VS ['PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuilding(s)', 'NumberofBuildings', 'Anciennete']
- TotalGHGEmissions VS ['PropertyGFATotal', 'PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'measure floors']
- TotalGHGEmissions VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']
- TotalGHGEmissions VS ['PropertyGFATotal', 'Anciennete', 'measure floors']
- TotalGHGEmissions VS ['PropertyGFATotal', 'PropertyGFABuilding(s)', 'Anciennete']
- TotalGHGEmissions VS ['PropertyGFATotal', 'Anciennete', 'measure floors']
- log TotalGHGEmissions VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- log TotalGHGEmissions VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'log PropertyGFABuilding(s)', 'NumberofBuildings', 'Anciennete']
- log TotalGHGEmissions VS ['log PropertyGFATotal', 'log PropertyGFAParking', 'NumberofBuildings', 'Anciennete', 'log measure floors']
- log TotalGHGEmissions VS ['log PropertyGFATotal', 'Anciennete', 'log measure floors']
- log TotalGHGEmissions VS ['log PropertyGFATotal', 'log PropertyGFABuilding(s)', 'Anciennete']

Estimateurs testés:

- Dummy
- Linear
- Ridge
- Lasso
- Elastic Net
- Random Forest

Émissions, 5 variables

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	342.27	-0.00	295.87	-0.00
Linear	0.00	355.26	0.23	240.68	0.23
Ridge	0.35	312.33	0.14	256.25	0.26
Lasso	3.06	218.29	0.18	278.48	0.24
Elastic Net	0.06	239.26	0.24	274.22	0.23
RF	8.08	218.16	0.63	134.17	0.79

Target: TotalGHGEmissions

Features: [log PropertyGFATotal, log PropertyGFAParking, NumberofBuildings, Anciennete, log measure floors]

- Le passage au logarithme augmente les r2, mais seulement un estimateur (à nouveau Random Forest) atteint une valeur r2>0.7 pour le jeu de test.
- Le passage au logarithme de la variable target se révèle indispensable pour obtenir une performance juste.

- r2 moins élevés

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.32	-0.02	1.37	-0.00
Linear	0.00	1.06	0.48	1.04	0.39
Ridge	0.38	1.00	0.47	1.05	0.40
Lasso	0.44	1.05	0.42	1.04	0.41
Elastic Net	0.07	0.98	0.33	1.06	0.42
RF	8.98	0.72	**0.70**	0.52	**0.86**

Target: log TotalGHGEmissions

Features: [log PropertyGFATotal, log PropertyGFAParking, NumberofBuildings, Anciennete, log measure floors]

Émissions, 3 variables

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	262.53	-0.01	316.05	-0.00
Linear	0.00	296.78	0.17	260.09	0.25
Ridge	0.34	201.42	0.22	282.12	0.23
Lasso	6.01	301.11	0.20	258.44	0.24
Elastic Net	0.06	245.43	0.17	273.11	0.24
RF	7.39	251.81	0.38	108.36	0.87

Target: TotalGHGEmissions

Features: [log PropertyGFATotal, Anciennete, log measure floors]

- Comme dans le cas précèdent, le passage au logarithme augmente les r2, mais seulement un estimateur atteint une valeur r2>0.7 pour le jeu d'entraînement.
- Le passage au logarithme de la variable target se révèle indispensable pour obtenir une performance discrète.

- r2 moins élevés

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.40	-0.02	1.35	-0.00
Linear	0.00	1.04	0.47	1.05	0.39
Ridge	0.37	1.10	0.38	1.04	0.42
Lasso	0.40	1.05	0.37	1.04	0.42
Elastic Net	0.07	0.99	0.46	1.06	0.40
RF	8.29	0.74	**0.71**	0.46	**0.88**

Target: log TotalGHGEmissions

Features: [log PropertyGFATotal, Anciennete, log measure floors]

Modèle retenu pour la prédiction des émissions

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.40	-0.02	1.35	-0.00
Linear	0.00	1.04	0.47	1.05	0.39
Ridge	0.37	1.10	0.38	1.04	0.42
Lasso	0.40	1.05	0.37	1.04	0.42
Elastic Net	0.07	0.99	0.46	1.06	0.40
RF	8.29	0.74	**0.71**	0.46	**0.88**

Pouvoir prédictif ~ 70%

- Target: log TotalGHGEmissions
- Features: [log PropertyGFATotal, Anciennete, log measure floors]

4 - Intérêt de l'Energy Star Score

Energy Star Score et consommation énergétique

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.21	-0.00	1.11	-0.00
Linear	0.00	0.69	**0.65**	0.64	**0.68**
Ridge	0.35	0.63	**0.69**	0.66	**0.67**
Lasso	0.36	0.56	**0.71**	0.67	**0.66**
Elastic Net	0.07	0.59	**0.71**	0.67	**0.66**
RF	7.33	0.48	**0.81**	0.30	**0.93**

Target: log SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, Anciennete, log measure floors]

- L'ajout de la feature ENERGYSTARScore parmi les features améliore la performance de tous les estimateurs.
- Le modèle le plus performant, Random Forest, voit pas sa performance s'améliorer du 7% sur le jeu de test.
- De plus, le décalage entre les r2 des jeux de test et d'entraînement se réduit aussi : dès 1.2% à 0.9%.

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.06	-0.00	1.11	-0.00
Linear	0.00	0.51	**0.78**	0.43	**0.85**
Ridge	0.33	0.50	**0.76**	0.43	**0.85**
Lasso	0.37	0.54	**0.77**	0.42	**0.85**
Elastic Net	0.06	0.53	**0.78**	0.42	**0.85**
RF	6.95	0.35	**0.88**	0.19	**0.97**

Target: log SiteEnergyUse(kBtu)

Features: [log PropertyGFATotal, Anciennete, log measure floors, ENERGYSTARScore]

Energy Star Score et émissions de gaz à effet de serre

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.40	-0.02	1.35	-0.00
Linear	0.00	1.04	0.47	1.05	0.39
Ridge	0.37	1.10	0.38	1.04	0.42
Lasso	0.40	1.05	0.37	1.04	0.42
Elastic Net	0.07	0.99	0.46	1.06	0.40
RF	8.29	0.74	**0.71**	0.46	**0.88**

Target: log TotalGHGEmissions

Features: [log PropertyGFATotal, Anciennete, log measure floors]

- L'ajout de la variable ENERGYSTARScore parmi les features améliore la performance des estimateurs Linear, Ridge, Lasso et Elastic Net.
- Cependant, le modèle le plus performant, Random Forest, ne voit pas sa performance s'améliorer significativement.

name	runtime	rmse test	r2 test	rmse train	r2 train
Dummy	0.00	1.44	-0.00	1.31	-0.00
Linear	0.00	0.83	0.63	0.86	0.58
Ridge	0.33	0.91	0.60	0.85	0.59
Lasso	0.35	0.92	0.55	0.84	0.60
Elastic Net	0.06	0.95	0.53	0.83	0.61
RF	6.96	0.69	**0.70**	0.44	**0.89**

Target: log TotalGHGEmissions

Features: [log PropertyGFATotal, Anciennete, log measure floors, ENERGYSTARScore]

Perspectives

On pourrait élargir l'analyse en tenant compte de :

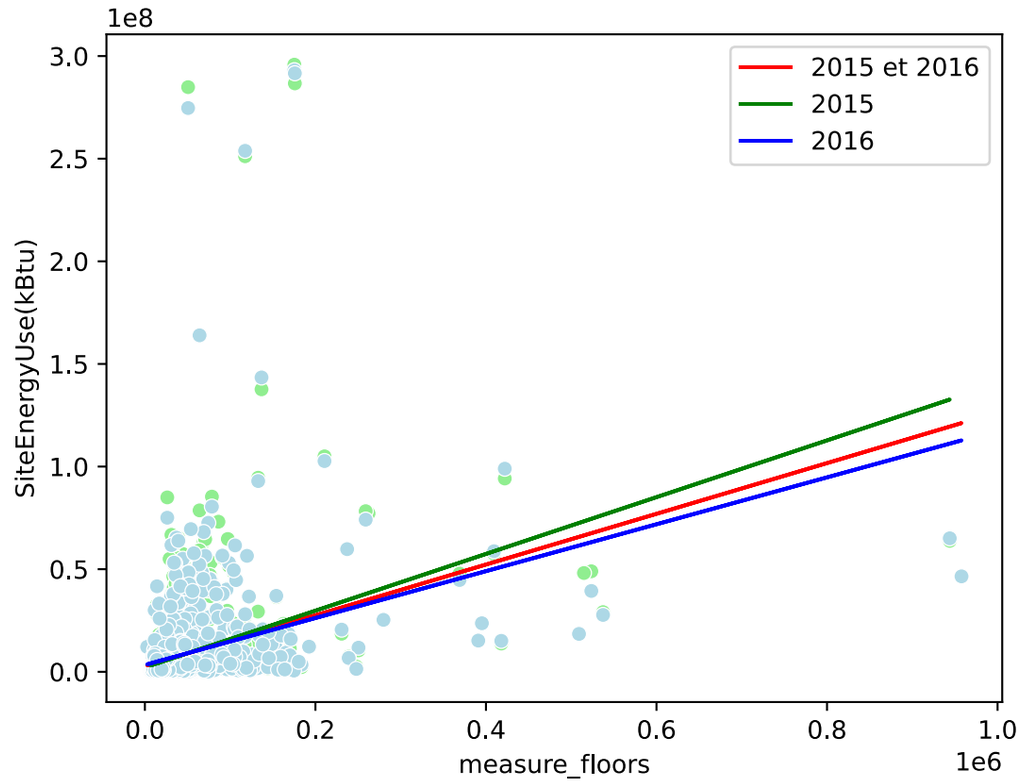
- L'utilisation du bâtiment, qui détermine ses tendances de consommation énergétique.
- Le climat de l'année en question (températures alignées ou pas avec la moyenne).

Merci

Stratégie finalisée à limiter l'overfitting

Approche choisie

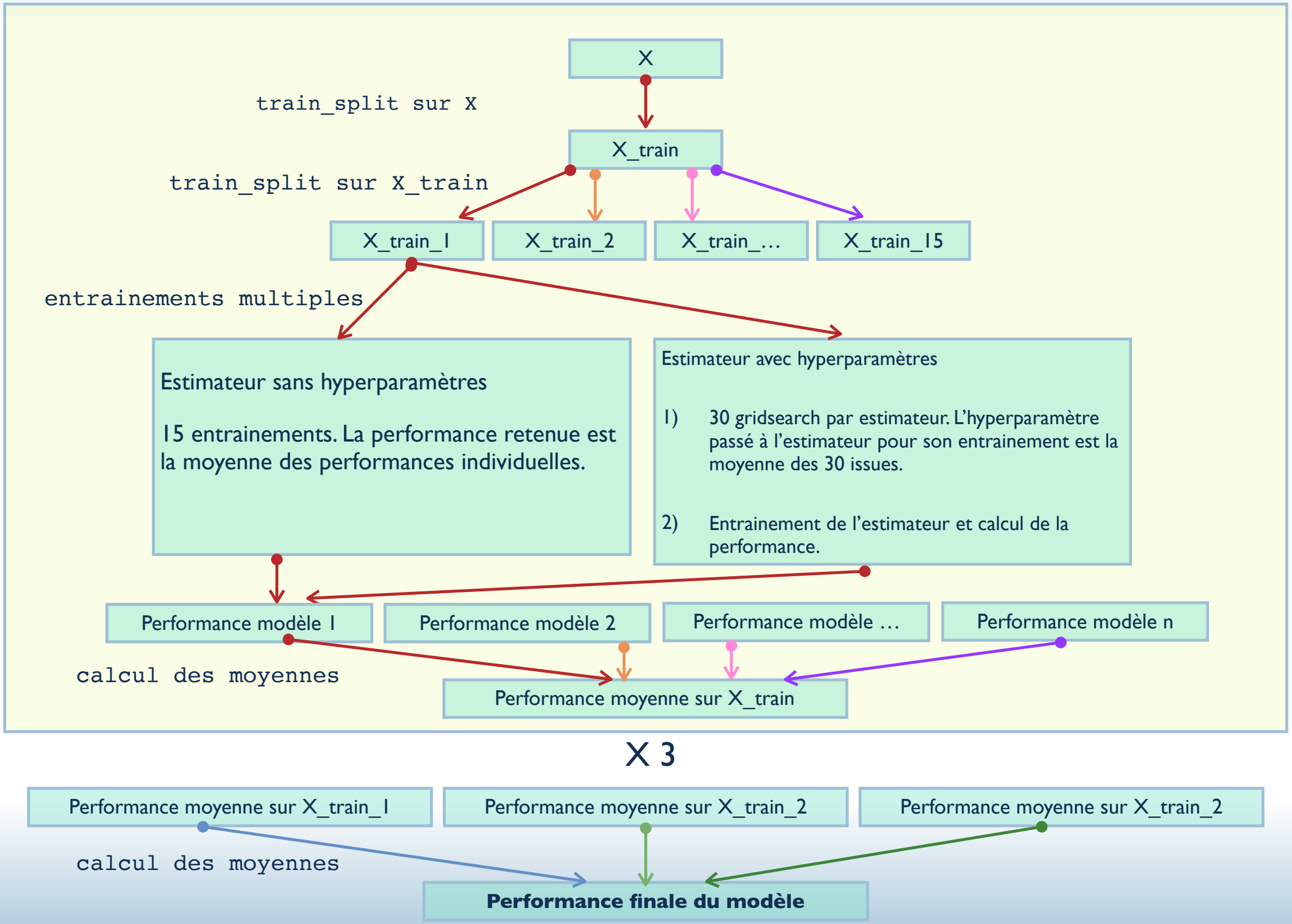
- Entraînements multiples sur différentes sections de X
- Moyennes



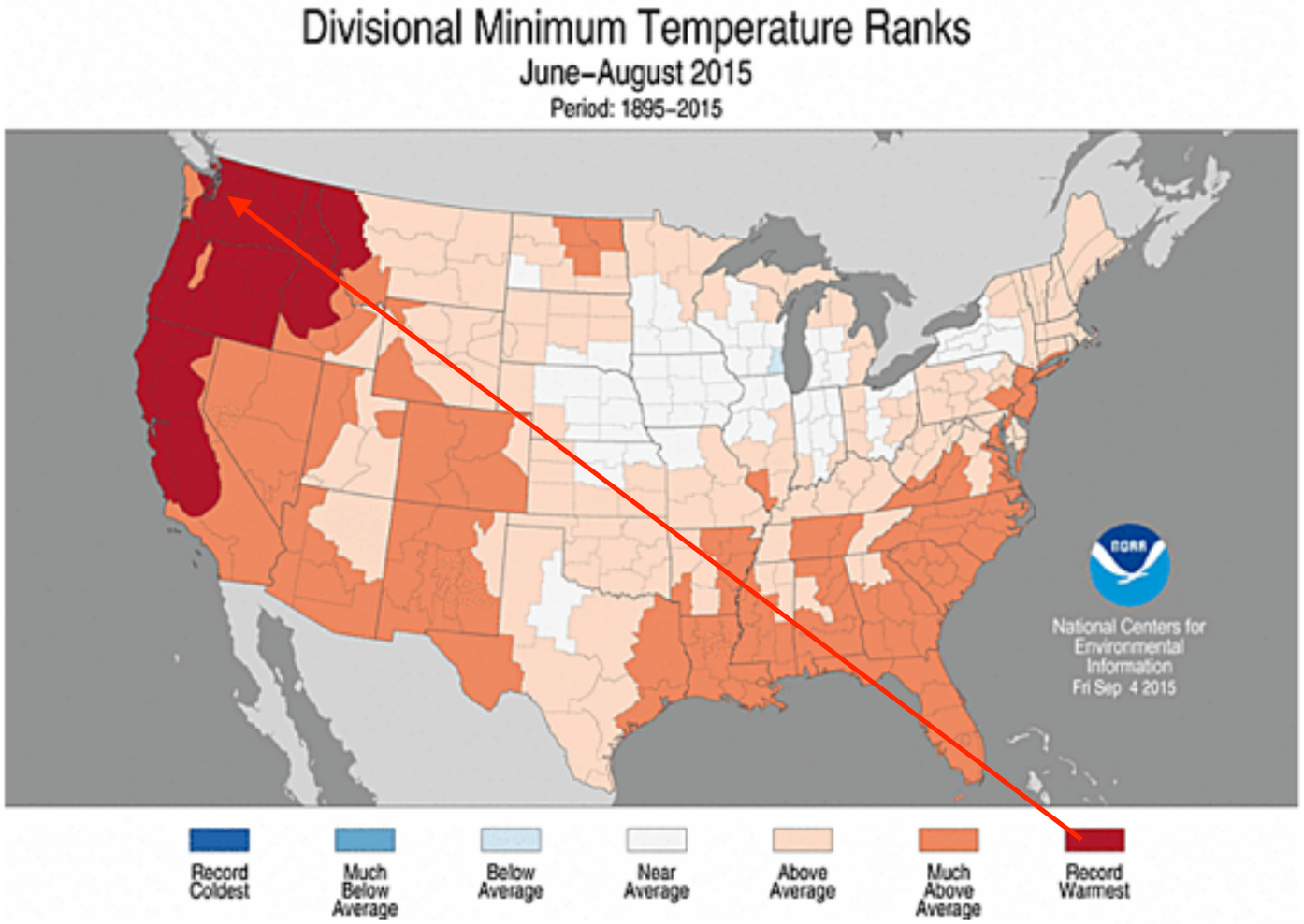
- Gridsearch multiples
- Moyennes

```
END best_alpha_rmse_1 and best_alpha_r2_1 (0, 'Ridge ', 3.1622776601683795e-05, 3.1622776601683795e-05)
END best_alpha_rmse_1 and best_alpha_r2_1 (1, 'Ridge ', 100.0, 100.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (2, 'Ridge ', 316.22776601683796, 316.22776601683796)
END best_alpha_rmse_1 and best_alpha_r2_1 (3, 'Ridge ', 100.0, 100.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (4, 'Ridge ', 316.22776601683796, 316.22776601683796)
END best_alpha_rmse_1 and best_alpha_r2_1 (0, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (1, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (2, 'Lasso ', 10.0, 10.0)
END best_alpha_rmse_1 and best_alpha_r2_1 (3, 'Lasso ', 31.622776601683793, 31.622776601683793)
END best_alpha_rmse_1 and best_alpha_r2_1 (4, 'Lasso ', 31.622776601683793, 31.622776601683793)
```

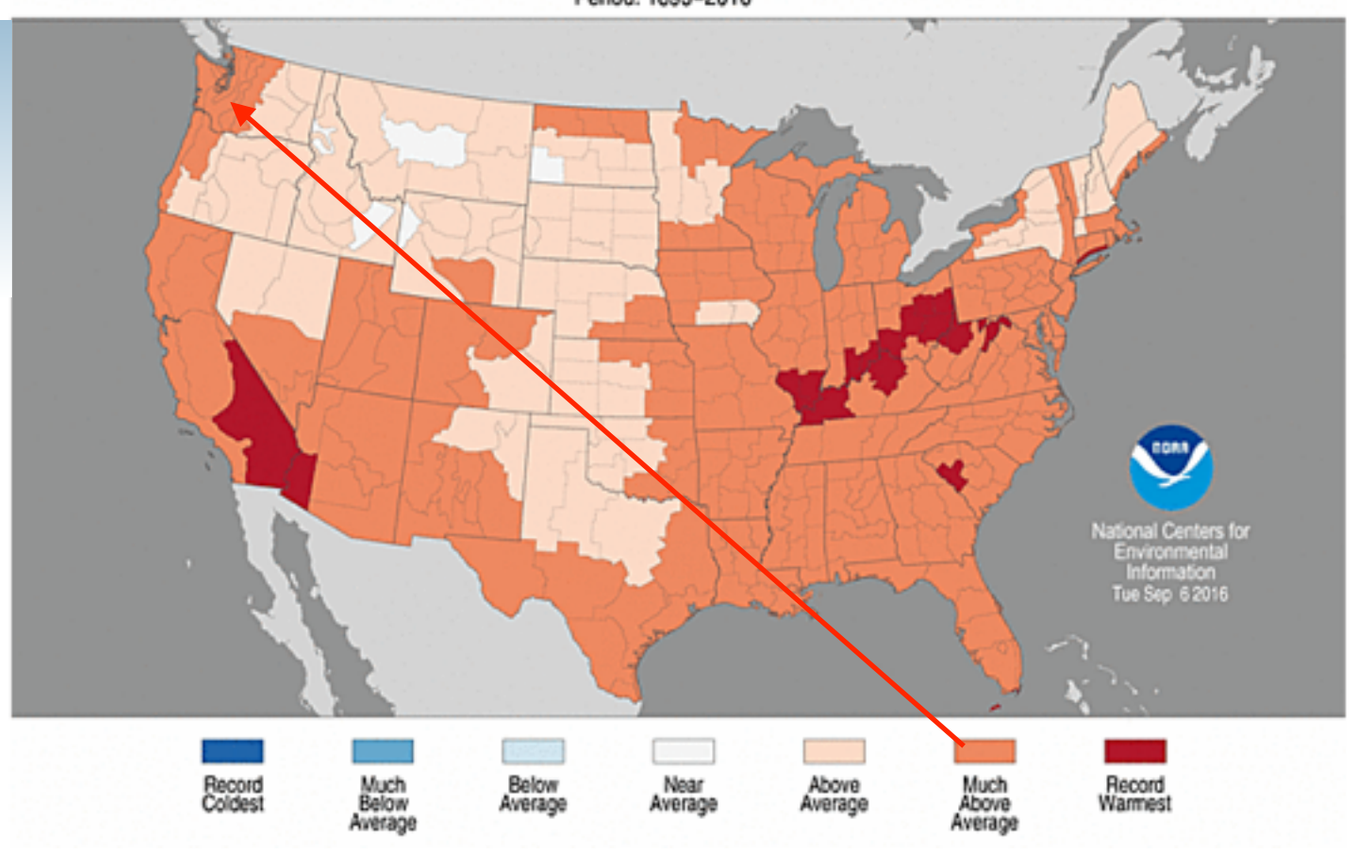
Stratégie finalisée à limiter l'overfitting



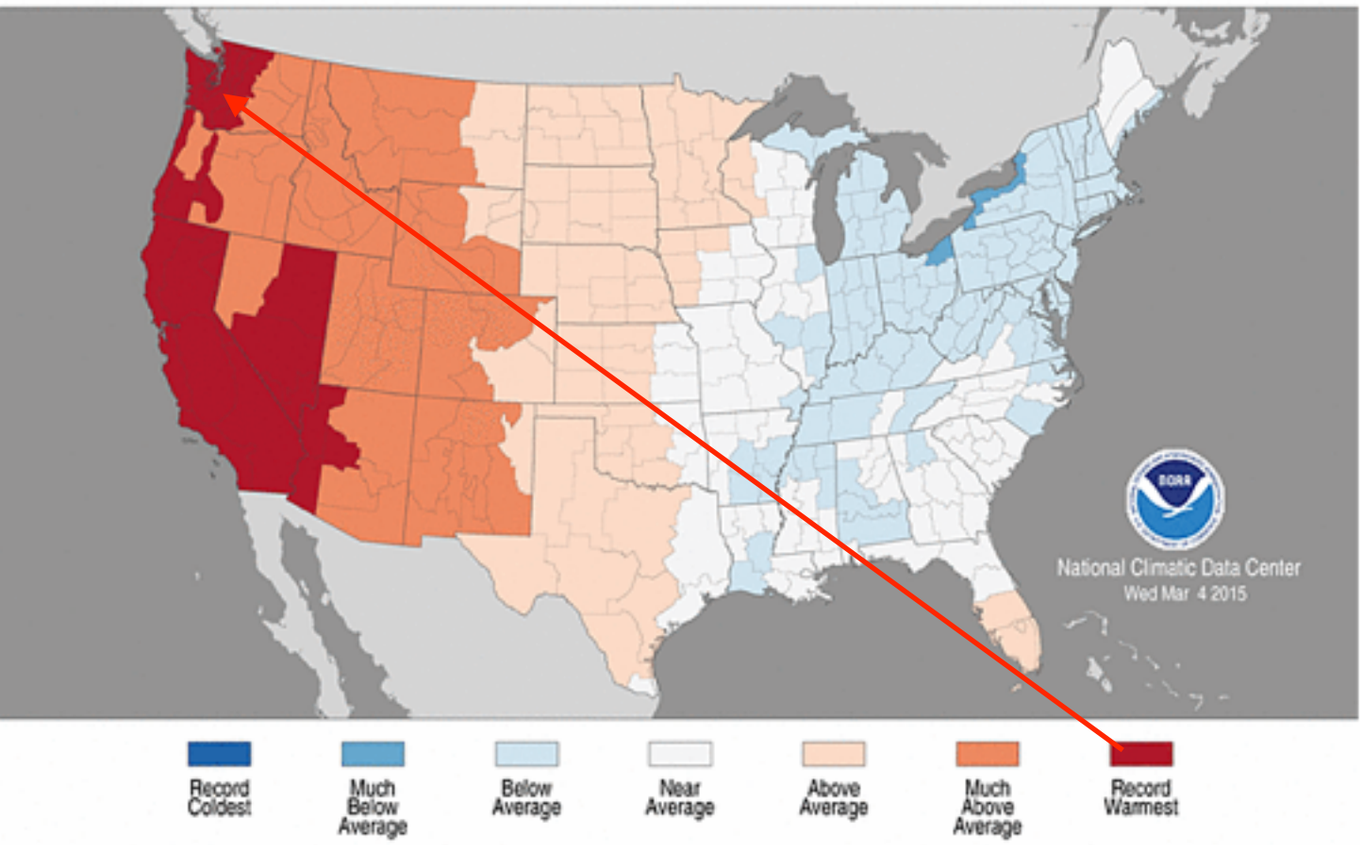
Été 2015



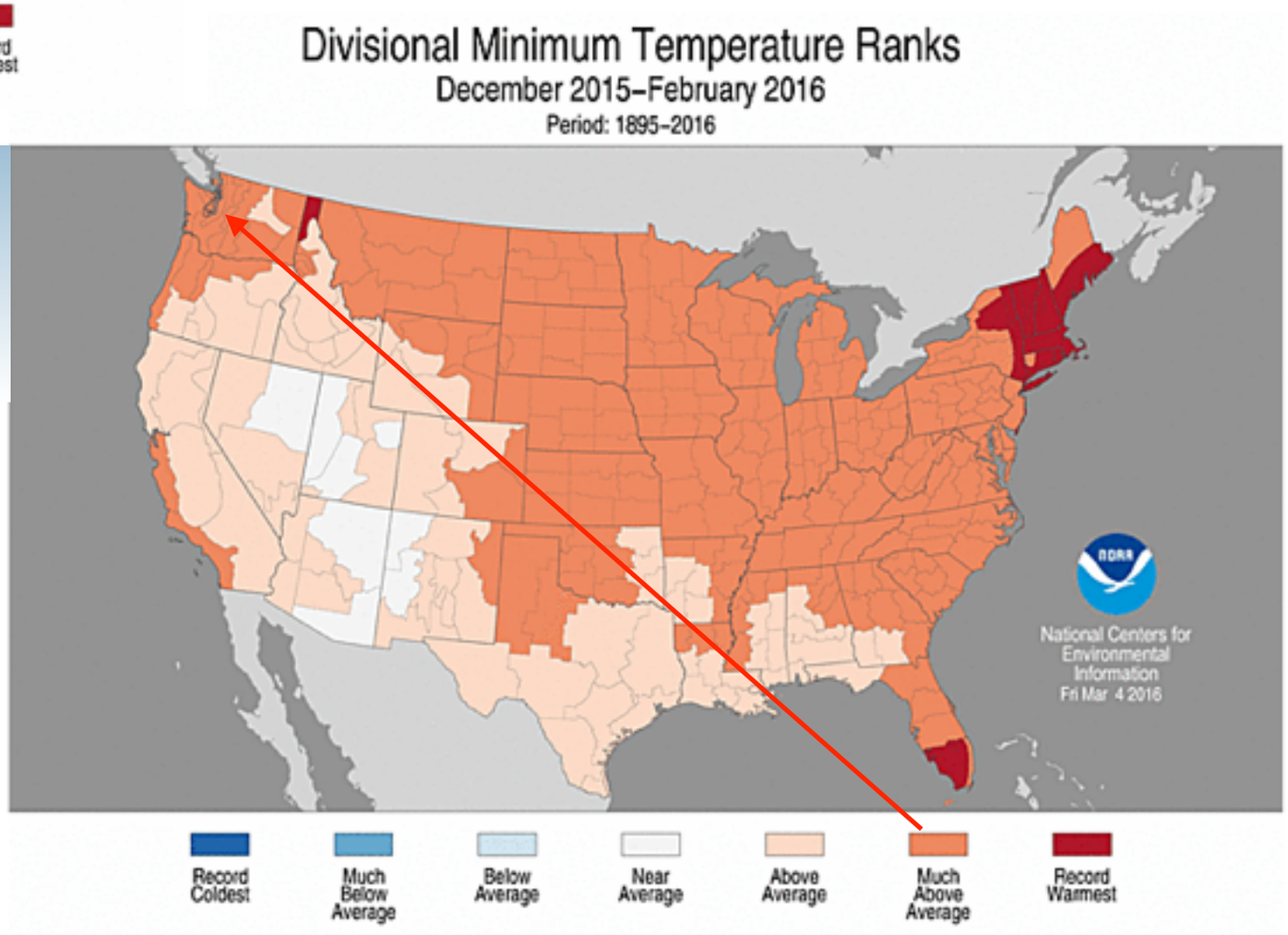
Été 2016



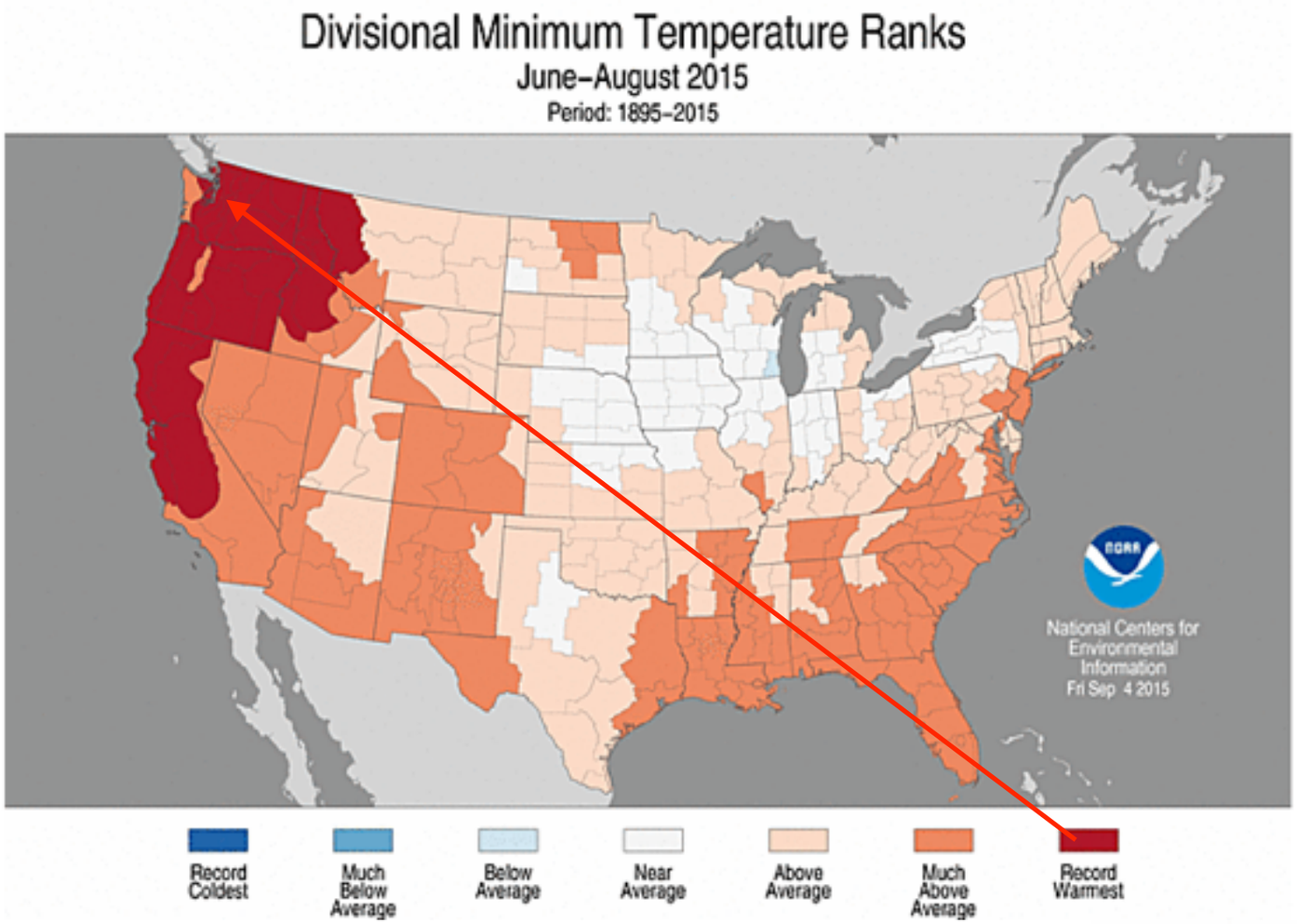
Hiver 2015



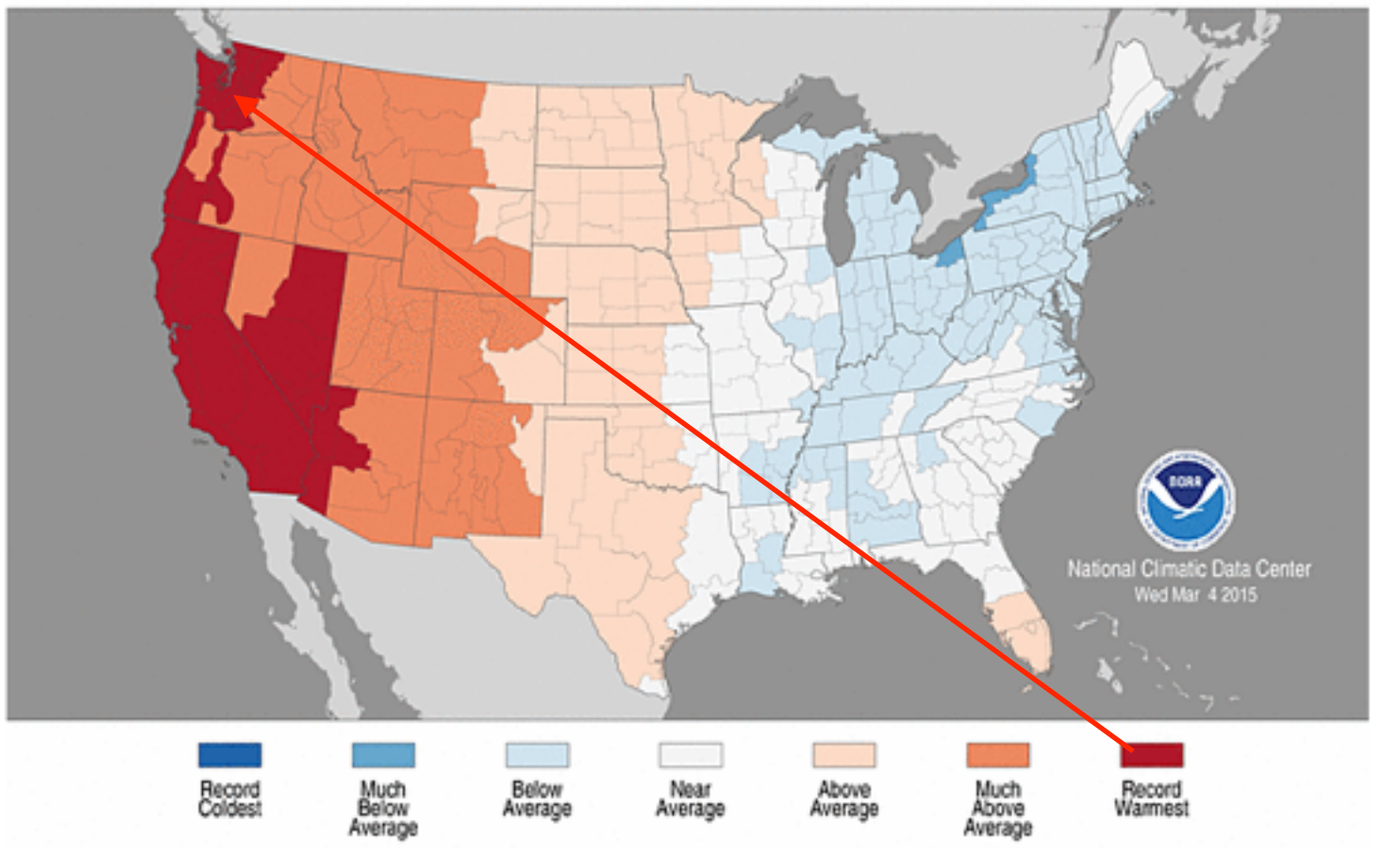
Hiver 2016



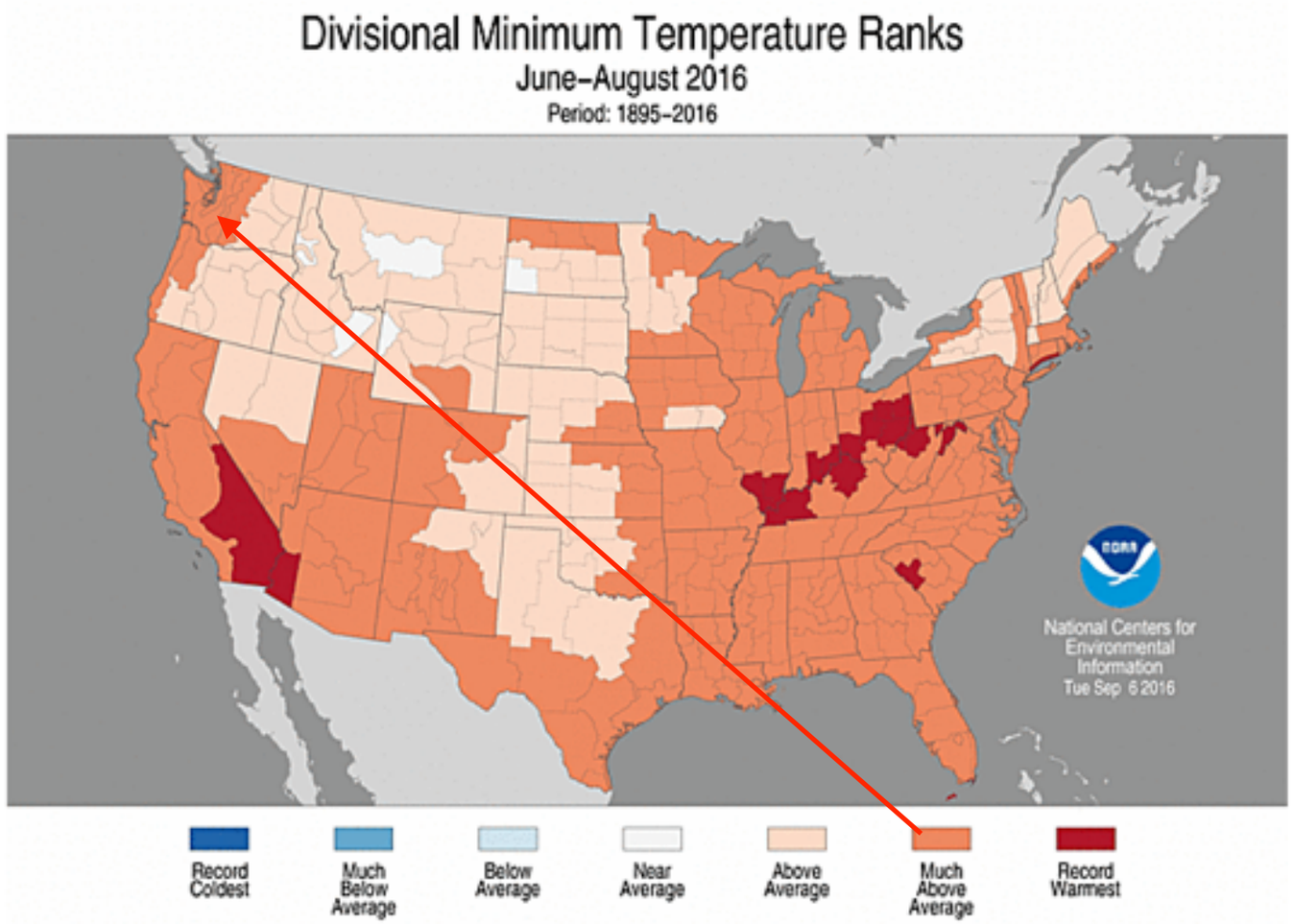
Été 2015



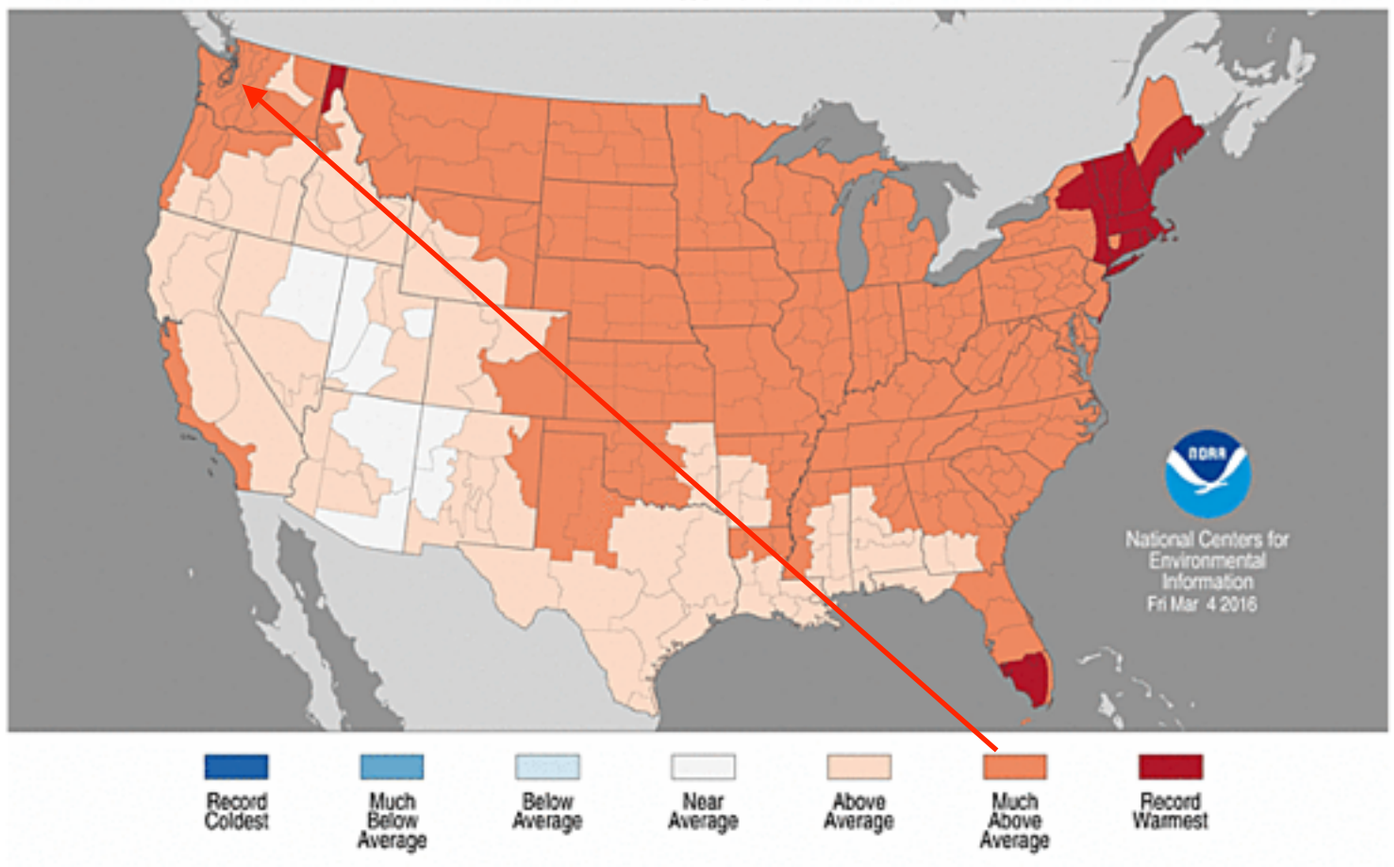
Hiver 2015



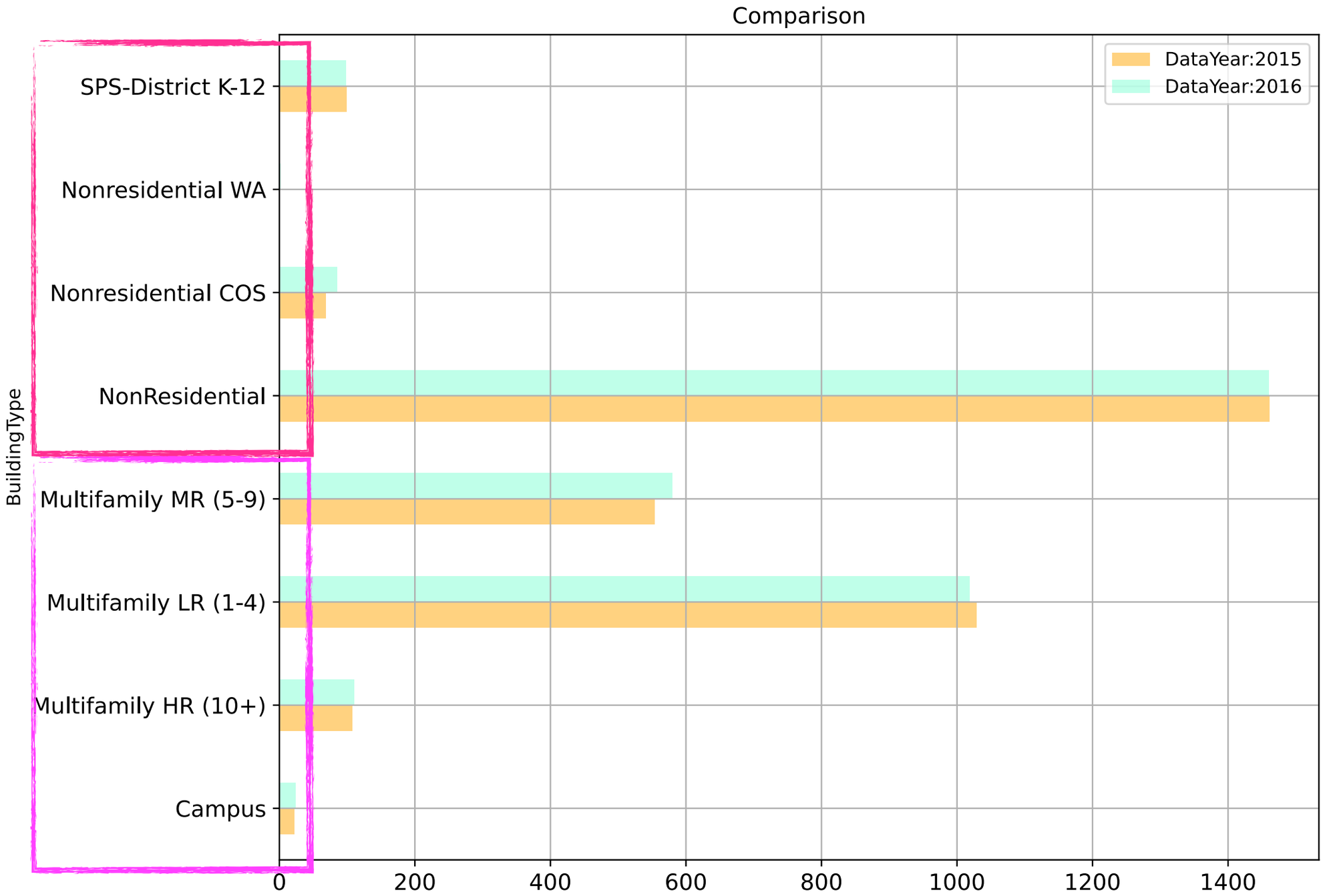
Été 2016



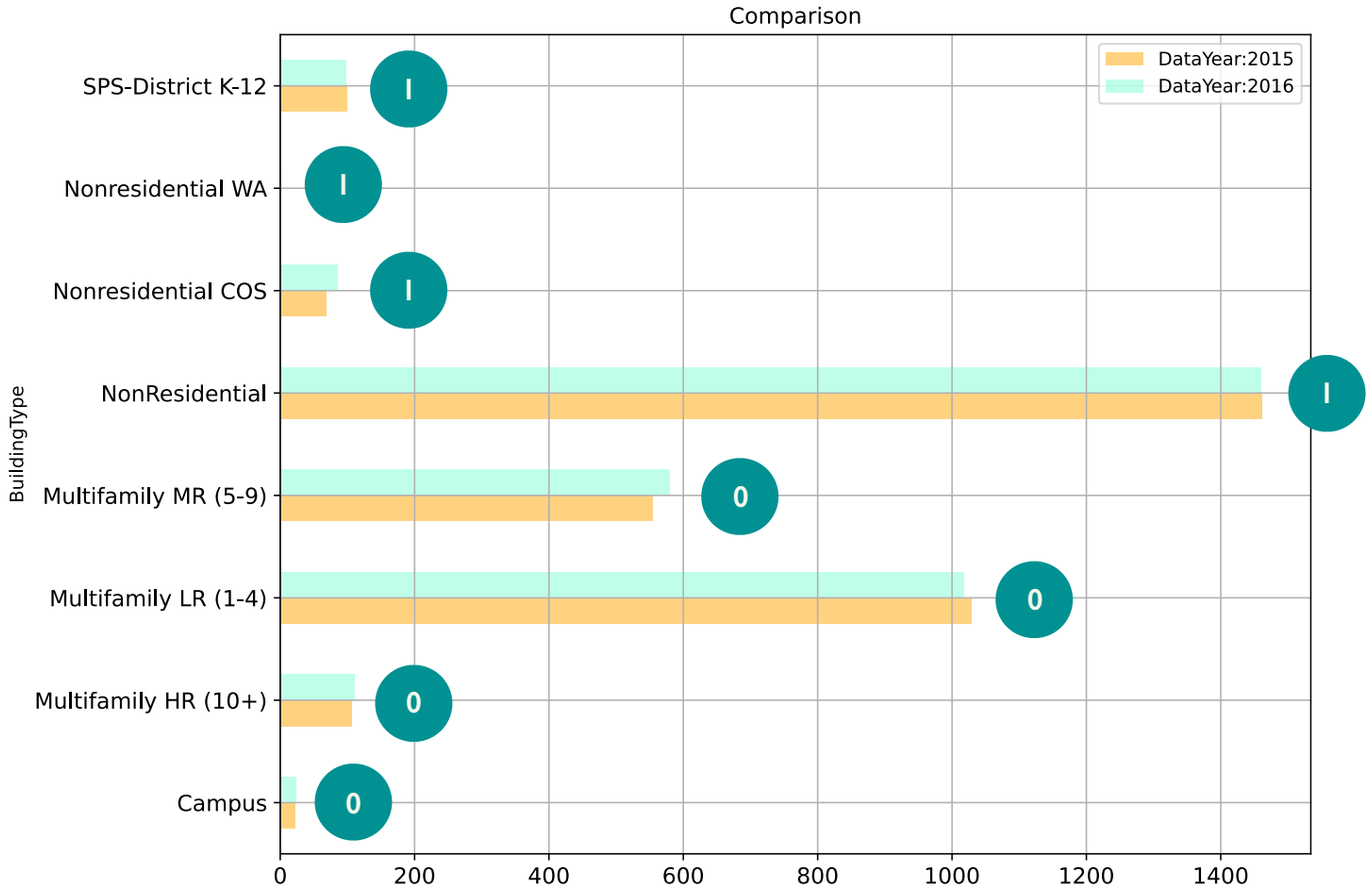
Hiver 2016



Type de résidence (BuildingType)



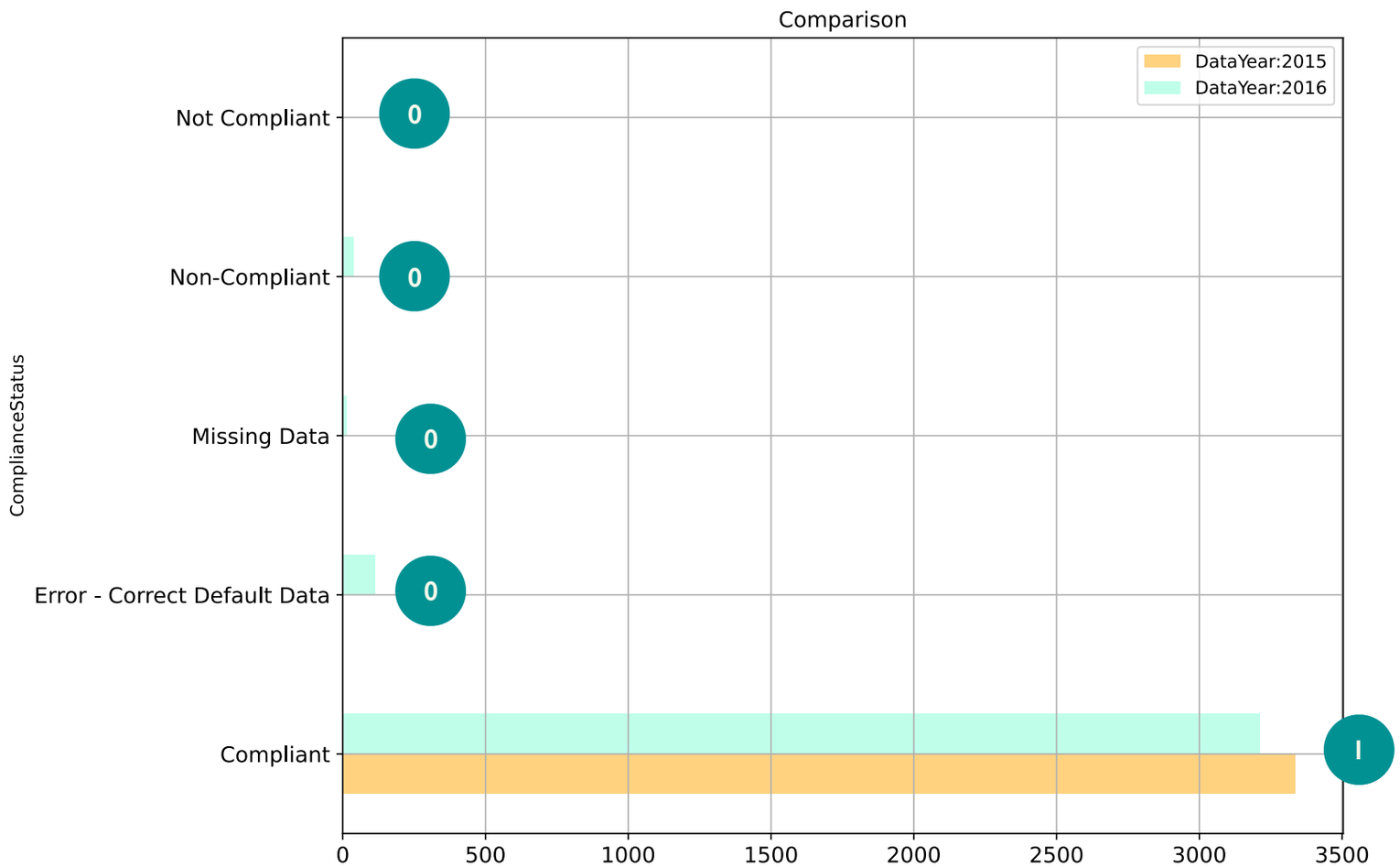
Sélection des bâtiments non résidentiels



On associe la valeur 1 aux bâtiments non résidentiels et 0 aux bâtiment résidentiels par l'ajout d'une colonne nommée "NonRes". On garde seulement les données relatives aux bâtiments non résidentiels, ce qui réduit les jeux des données à:

- 2015: 3340 → **1628** lignes x 48 colonnes
- 2016: 3376 → **1644** lignes x 46 colonnes

Ajout de colonnes: 1 - conformités aux standards

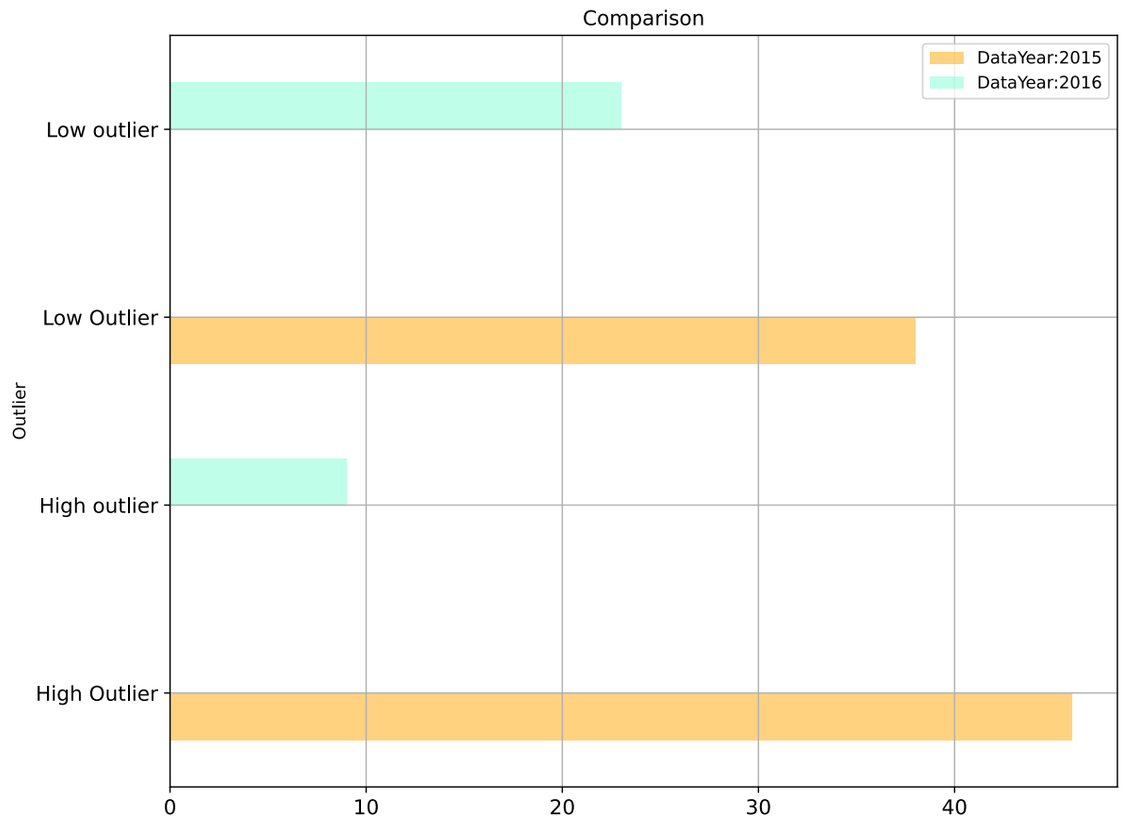


On associe la valeur 1 aux propriétés conformes aux standards et 0 à tous les autres dans la nouvelle colonne “Compl”.

- 2015: toutes les propriétés sont marqués conformes, elle sont donc 1628.
- 2016: Sur 1644 propriétés, 1524 sont marquées conformes.

Ajout de colonnes 2 - outliers

Colonne “Outlier”: notation non homogène.



Dans la nouvelle colonne “Outlier_ok”, on associe la lettre “L” aux outliers bas et “H” aux outliers hauts pour rendre uniforme la notation.