Prédiction du Taux de Grippe en France

Challenge Kaggle - Machine Learning Appliqué

3 janvier 2026

1 Contexte et Objectif

Ce projet s'inscrit dans le cadre d'une compétition Kaggle visant à prédire le **taux de grippe hebdomadaire** pour les 22 régions françaises métropolitaines. L'objectif principal est de développer un modèle prédictif robuste capable d'estimer le nombre de consultations pour syndrome grippal à partir de données historiques couvrant la période 2004–2011, puis de généraliser ces apprentissages pour prédire l'ensemble de l'année 2012.

Ce défi présente plusieurs enjeux majeurs. D'une part, la grippe saisonnière constitue un problème de santé publique significatif, avec des pics d'incidence pouvant saturer les systèmes de soins. D'autre part, la capacité à anticiper ces pics permettrait une meilleure allocation des ressources hospitalières, une planification plus efficace des campagnes de vaccination et une communication préventive auprès du grand public.

Nous cherchons à produire des prédictions fiables, maintenables et interprétables afin d'appuyer des décisions de santé publique (surveillance épidémiologique, allocation de ressources médicales) et d'évaluer la robustesse des signaux exogènes tels que les conditions météorologiques et les recherches web. Notre approche privilégie délibérément la simplicité et la généralisation afin d'éviter des modèles trop complexes qui seraient excessivement sensibles aux anomalies historiques, notamment l'année exceptionnelle 2009 marquée par la pandémie H1N1.

Métrique d'évaluation : La performance est mesurée par le RMSE (Root Mean Squared Error), métrique standard utilisée pour comparer les soumissions sur la plateforme Kaggle. Cette mesure quantifie l'écart quadratique moyen entre les valeurs prédites et les observations réelles, pénalisant davantage les erreurs importantes que les petites déviations.

Données disponibles : Nous disposons de quatre sources complémentaires. Les **données SYNOP** regroupent les relevés météorologiques horaires de 67 stations réparties sur le territoire métropolitain (température, humidité, précipitations, vent). Les **données Google Trends** fournissent un indice mensuel de popularité des recherches liées à la grippe, décliné par région. Les **données démographiques INSEE** apportent la population annuelle par région et tranche d'âge, indispensable pour calculer les taux d'incidence. Enfin, l'**historique des taux de grippe** (2004–2011), issu du réseau Sentinelles, constitue notre variable cible.

La richesse de ces sources de données offre une opportunité de modélisation multi-factorielle, mais impose également des défis d'intégration liés aux différences de granularité temporelle et spatiale.

2 Préparation des Données

La principale difficulté réside dans l'hétérogénéité temporelle et spatiale des différentes sources de données. En effet, les données météorologiques sont horaires, Google Trends est mensuel, la population est annuelle, tandis que notre cible est hebdomadaire. De plus, le découpage géographique diffère selon les sources (stations ponctuelles vs régions administratives). Pour résoudre ces problèmes d'alignement, nous avons développé un pipeline de fusion rigoureux en trois étapes successives.

2.1 Agrégation Météorologique

Les données SYNOP horaires ont été agrégées à la semaine ISO par région. Pour la **température**, nous calculons la moyenne, le minimum, le maximum et l'écart-type hebdomadaires. L'**humidité** est résumée par sa moyenne, son minimum et son maximum. Les **précipitations** sont caractérisées par leur somme, leur moyenne et leur valeur maximale sur la semaine. Enfin, le **vent** est décrit par sa vitesse moyenne et sa vitesse maximale.

Pour l'agrégation, chaque station a été mappée à une région administrative en utilisant ses coordonnées géographiques (latitude, longitude). Les statistiques hebdomadaires sont ensuite calculées par combinaison non pondérée des stations présentes dans chaque région. Lorsque la densité de stations est hétérogène entre régions, nous réalisons des contrôles qualitatifs pour éviter les biais systématiques (par exemple, une surreprésentation des stations côtières par rapport aux stations intérieures dans certaines régions).

Les agrégations sont conçues pour être robustes aux valeurs extrêmes et aux variations saisonnières, afin de ne pas biaiser les moyennes régionales par quelques observations aberrantes. Nous utilisons notamment la médiane pour les imputations et l'écart-type pour capturer la variabilité intra-semaine.

Défi Franche-Comté : Cette région ne possède aucune station SYNOP sur son territoire, ce qui constitue un défi méthodologique. Pour y remédier, nous avons assigné la station frontalière de Bâle-Mulhouse (située en Alsace mais proche géographiquement) en nous appuyant sur la proximité géographique et la similarité climatologique entre ces deux régions. Ce choix a été documenté et nous avons testé la sensibilité des résultats à cette substitution, confirmant un impact limité sur les performances globales.

2.2 Fusion Google Trends

Les données Google Trends, initialement disponibles à la granularité mensuelle, ont été converties en données hebdomadaires via une correspondance entre semaine ISO et mois calendaire. Pour chaque semaine, nous utilisons le jeudi comme date d'ancrage pour déterminer le mois correspondant, suivant la convention ISO 8601.

Avant la fusion avec les autres sources, chaque série Google Trends régionale a été normalisée par son maximum historique. Cette normalisation permet de réduire l'effet d'échelle entre régions (certaines régions ont naturellement plus de recherches que d'autres en raison de leur population connectée) et d'atténuer la variabilité liée à l'évolution de la popularité d'Internet au fil des années.

Nous avons également testé une version lissée des données Google Trends (moyenne mobile sur 3 semaines) pour évaluer l'impact du bruit à court terme sur les performances du modèle. Les résultats montrent que le signal brut reste plus informatif, suggérant que les variations rapides des recherches contiennent de l'information prédictive utile.

2.3 Interpolation Population

Les données démographiques INSEE, publiées à fréquence annuelle, ont été interpolées à la granularité hebdomadaire via une **spline cubique**. Cette méthode mathématique permet d'obtenir une évolution progressive et réaliste de la population au cours de l'année, évitant les discontinuités brutales au passage d'une année à l'autre.

La spline a été soigneusement paramétrée pour garantir une dérivée première continue aux frontières annuelles, assurant ainsi des transitions fluides entre les années. Cette contrainte évite les fluctuations artificielles qui pourraient apparaître en début ou fin d'année civile. L'interpolation ainsi obtenue est suffisante pour capturer les grandes tendances démographiques (croissance,

vieillissement) sans introduire de signal hebdomadaire artificiel qui pourrait parasiter l'apprentissage du modèle.

2.4 Traitement des Valeurs Manquantes

Variable	Méthode	Justification
Météo	Médiane	Robuste aux outliers
Google Trends	0	Absence = pas d'intérêt
Lags	Médiane globale	Évite NaN en début de série

Lorsque plusieurs variables manquent simultanément pour une région/semaine, nous évaluons la consistance spatiale (imputation par voisinage régional) avant d'imputer globalement. Les imputations sont conservatrices afin de limiter l'introduction de signal artificiel.

3 Modélisation

3.1 Évolution Itérative

Le développement du modèle a suivi une démarche itérative rigoureuse, avec pas moins de 14 versions successives. Nous sommes partis d'un modèle initialement complexe intégrant 43 features (toutes les variables météorologiques, démographiques et comportementales disponibles) pour aboutir à un modèle épuré et régularisé ne conservant que 15 features soigneusement sélectionnées.

Le constat principal, contre-intuitif au premier abord, est le suivant : **plus le modèle était simple, meilleur était le score obtenu sur Kaggle**. Ce résultat met en évidence un problème d'overfitting caractéristique des situations où le nombre d'observations est limité par rapport au nombre de prédicteurs potentiels.

La démarche d'itération a été guidée par des tests empiriques systématiques. L'**ablation de features** a consisté à retirer successivement des groupes de variables pour mesurer leur contribution réelle aux performances. La **régularisation croissante** s'est traduite par une augmentation progressive des pénalités L2 et une réduction de la profondeur des arbres. Enfin, des **tests de robustesse** ont permis d'évaluer la sensibilité du modèle à l'inclusion ou l'exclusion de l'année 2009, marquée par la pandémie H1N1.

À chaque version, nous avons conservé des métriques de validation détaillées ainsi que des diagnostics complémentaires (importance des features, distribution des résidus, courbes d'apprentissage) pour comparer objectivement la stabilité et la capacité de généralisation des différentes configurations.

Version	Features	Val RMSE	Kaggle
V1 (baseline)	13	63.78	–
V2 (enrichi)	43	56.76	>100
V7 (mix)	18	58.17	95
V10 (anti-overfit)	18	67.58	91
V12 (final)	**15**	**69.07**	**88**

3.2 Modèle Final (V12)

Algorithme : Nous avons retenu le `CatBoostRegressor`, un algorithme de gradient boosting développé par Yandex, reconnu pour sa robustesse et sa capacité à gérer efficacement les données

tabulaires. Cet algorithme présente plusieurs avantages dans notre contexte : gestion native des valeurs manquantes, régularisation intégrée et excellentes performances sans nécessiter de prétraitement extensif des features.

La prédiction s'effectue de manière récursive pour l'année 2012, chaque semaine utilisant les prédictions précédentes comme features de lag.

Hyperparamètres anti-overfitting : Les hyperparamètres ont été soigneusement ajustés pour maximiser la généralisation plutôt que la performance sur l'ensemble d'entraînement. Nous utilisons des arbres peu profonds (`depth=4`), une forte régularisation L2 (`l2_leaf_reg=12`), un nombre minimal d'observations par feuille élevé (`min_data_in_leaf=60`) pour favoriser la généralisation, un taux d'apprentissage lent (`learning_rate=0.03`) et 800 itérations.

Les 15 features sélectionnées :

Catégorie	Features
Temporel	`week_num`, `sin_k`, `cos_k`, `is_flu_season`
Lags	`taux_lag1`, `taux_lag2`, `taux_lag3`, `taux_ma3`
Google	`google_grippe`, `google_lag1`
Historique	`rw_mean`, `rw_median`, `w_mean`, `r_mean`
Interaction	`google_x_rw`

3.3 Stratégie de Validation

Pour simuler la contrainte temporelle réelle (prédire 2012 avec données jusqu'à 2011), nous avons adopté une validation sur la **dernière année complète** (2011) plutôt qu'un split aléatoire 80/20. Concrètement, la validation retient les semaines 2004–2010 pour l'entraînement et 2011 pour la validation finale. Nous avons également testé une validation par fenêtres glissantes (rolling window) pour vérifier la stabilité temporelle des hyperparamètres et la sensibilité aux événements extrêmes.

3.4 Prédiction Récursive

Les features de lag (`taux_lag1`, etc.) nécessitent une stratégie récursive pour le test :

1. Prédire semaine t avec les vraies valeurs de $t-1, t-2, ...$
2. Utiliser la prédiction de t pour prédire $t+1$
3. Répéter pour toutes les semaines de 2012

Cette approche imite la contrainte réelle mais entraîne une accumulation d'erreurs qui augmente avec l'horizon ; nous analysons cet effet en regardant l'évolution du RMSE semaine par semaine sur la période test.

4 Résultats et Discussion

4.1 Analyse de l'Overfitting

L'écart entre validation (69) et Kaggle (88) révèle un overfitting résiduel dont nous avons identifié plusieurs causes. Premièrement, l'**année 2009** a créé des patterns atypiques en raison de la pandémie H1N1, avec une moyenne de 197 consultations contre 22 à 95 les autres années, ce qui biaise l'apprentissage. Deuxièmement, la **propagation d'erreur** inhérente à la prédiction récursive amplifie les écarts sur les 52 semaines de l'horizon de prédiction. Troisièmement, un **décalage de distribution** (distribution shift) est probable : l'année 2012 peut présenter des caractéristiques épidémiologiques légèrement différentes de la période d'entraînement.

En analysant les résidus par région et par semaine, on observe que certaines régions (forte densité urbaine) présentent une variabilité plus élevée, probablement liée à des comportements de

consultation et à l'utilisation d'Internet (Google Trends). L'influence de la météo est faible mais non nulle : elle améliore légèrement les prédictions pour les régions où la variabilité climatique coïncide avec des pics de grippe.

4.2 Features les Plus Importantes

1. `taux_lag1` (42%) - Le meilleur prédicteur est la valeur précédente
2. `rw_mean` (18%) - Moyenne historique région-semaine
3. `google_grippe` (12%) - Signal contemporain des recherches
4. `sin_k` / `cos_k` (8%) - Saisonnalité annuelle

Ces importances confirment l'intuition épidémiologique : la dynamique récente (lags) et la saisonnalité expliquent l'essentiel de la variance, tandis que les signaux exogènes (Google Trends, météo) jouent un rôle secondaire mais utile pour capturer des variations ponctuelles.

5 Conclusion

Les principaux enseignements tirés de ce travail sont les suivants.

Premièrement, **moins de features signifie meilleure généralisation** : dans un contexte de données limitées (8 années d'historique seulement), la réduction du nombre de prédicteurs s'est avérée plus efficace que l'enrichissement du modèle. Deuxièmement, **les lags temporels dominent** la prédiction : l'auto-corrélation temporelle de la série (la valeur de la semaine précédente) reste le prédicteur le plus puissant, confirmant la nature inertielle des épidémies. Troisièmement, **Google Trends apporte un signal complémentaire** précieux : les recherches web captent une information contemporaine que les lags seuls ne peuvent fournir, notamment en début de vague épidémique. Quatrièmement, **la météo a un impact limité** : une fois les autres features présentes, les variables météorologiques n'apportent qu'une contribution marginale, suggérant que leur effet est déjà capté indirectement par les autres variables. Cinquièmement, **la validation temporelle est cruciale** : un split aléatoire aurait gravement surestimé les performances réelles du modèle en créant une fuite d'information temporelle.

Améliorations futures : Plusieurs pistes prometteuses restent à explorer pour améliorer les performances et la robustesse du modèle. On pourrait envisager l'exclusion, la pondération ou une modélisation spécifique de l'année 2009 afin de réduire l'influence de cet événement exceptionnel sur l'entraînement. Un modèle hiérarchique, combinant une composante régionale et une composante globale, permettrait de capturer des dynamiques locales tout en partageant l'information entre régions. L'intégration de données épidémiologiques additionnelles (données détaillées du réseau Sentinelles, ventes d'antiviraux en pharmacie) enrichirait le signal prédictif. Enfin, l'exploration d'approches probabilistes ou de méthodes d'ensemble permettrait de mieux quantifier l'incertitude associée aux prédictions.

En résumé, le pipeline développé au cours de ce projet est robuste, reproductible et entièrement documenté. Il constitue une base solide et extensible pour des travaux futurs visant à améliorer la détection précoce et la prévision des épidémies saisonnières de grippe en France.

Au-delà de l'objectif immédiat de la compétition Kaggle, ce travail démontre l'intérêt d'une approche méthodologique rigoureuse combinant expertise métier (épidémiologie), ingénierie des données (fusion multi-sources) et bonnes pratiques de machine learning (validation temporelle, régularisation). Ces enseignements sont transférables à de nombreux autres problèmes de prévision dans le domaine de la santé publique.