

**100% ACTUAIRES &
100% DATA SCIENCE**

INSTITUT DES
ACTUAIRES



29 / NOV / 2019

Hôtel Marriott Rive Gauche
Paris 14ème

Accélérateurs pour la mise en place d'une couverture indiciaire des risques météo-sensibles



- 01 L'assurance indicielle**
Présentation et contexte de l'étude
- 02 Données météorologiques et support du risque**
Extractions et retraitements réalisés
- 03 La construction de l'indice**
Automatisation de la mise en place d'un réseau de neurones artificiels
- 04 Application à la création d'une couverture indicielle**
Création et application sur un cas pratique
- 05 Et après ?**
Conclusion et ouverture

01 - L'assurance indicielle

Les risques météorologiques en quelques chiffres



80%

Part de l'économie
météo-sensible¹



90%

Pertes économiques
liées aux intempéries
non couvertes par
l'assurance¹



66%

Part des secteurs
d'activité impactés
par le risque météo²



\$420 Mds

Montants soumis au
risque météo en France

¹ « L'assurance paramétrique, un marché prometteur », Les Echos

² « La gestion du risque météo en entreprise »

Quelques exemples d'activités météo-sensibles



Agro-alimentaire



Tourisme
(restauration, hôtellerie, etc.)



**Energies
renouvelables**



Gaz et Electricité

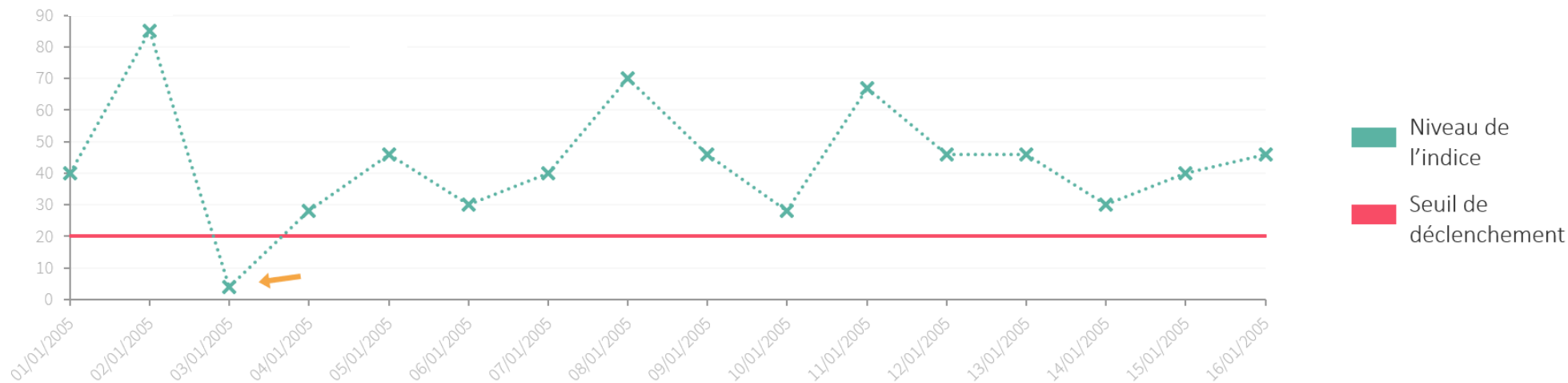


Logistique



Agriculture

Fonctionnement d'une couverture indiciaire



LE SUPPORT DU RISQUE

- Quantité numérique variant au cours du temps
- Le risque couvert correspond à une variation du support du risque

L'INDICE

- Correspond à une prédiction (estimation) du support du risque
- Effectuée à l'aide de variables météorologiques

LE SEUIL DE DÉCLENCHEMENT

- Peut-être fixe ou variable en fonction du temps
- Le franchissement du seuil par l'indice déclenche l'indemnisation

L'INDEMNISATION

- Automatique
- Peut correspondre à un montant forfaitaire ou être calculée à partir du niveau de l'indice

Avantages des couvertures indicielles étudiées

INDEMNISATION RAPIDE

Délais d'indemnisation de l'ordre
d'une semaine



ABSENCE D'ALÉA MORAL

Le niveau de l'indice ne peut être
influencé par le comportement de
l'assuré

PROVISIONS TECHNIQUES

Pas de provisions Dossier-Dossier,
IBNER, IBNYR

FRAIS DE GESTION FAIBLES

Pas de processus de déclaration,
aucune intervention d'expert
nécessaire

02 - Données météorologiques et support du risque

VARIABLES EXHAUSTIVES

42 variables : température, vitesse et direction du vent, précipitation, humidité, pression atmosphérique, irradiation solaire, etc.

DISPONIBILITÉ GLOBALE

Toutes les zones géographiques peuvent être couvertes, résolution spatiale de 0,5 latitude x 0,5 degrés longitude



EXTRACTION AUTOMATISÉE

Fenêtre géographique et temporelle paramétrable ; téléchargements asynchrones à l'aide de l'API de la NASA intégrée à **indexmethods**

HISTORIQUE IMPORTANT

Historique journalier pour l'ensemble des variables : 1983 – aujourd'hui (36 ans)

Illustration de l'incohérence de résolution spatiale

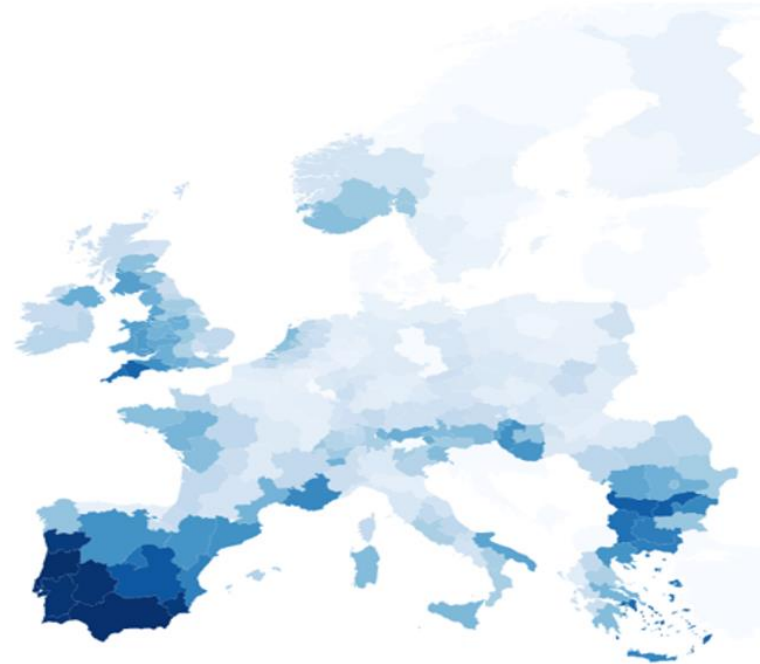
RÉSOLUTION SPATIALE DES
VARIABLES EXPLICATIVES

Maille de référence de l'étude
0,5 x 0,5 degrés



RÉSOLUTION SPATIALE
D'UN SUPPORT DU RISQUE
EXEMPLE

Zones administratives NUTS 2



Retraitement du support du risque

EXTRACTION DU SUPPORT DU RISQUE

La résolution spatiale
correspond au découpage
administratif NUTS 2 :
incohérence de résolution

01



FORMAT APPOSÉ À CELUI DES VARIABLES EXPLICATIVES

Chaque pixel se voit attribuer
une valeur si son centroïde est
dans une zone NUTS 2

02



LISSAGE PAR CONVOLUTIONS GAUSSIENNES

Un paramètre de lissage optimal
est déterminé tel que l'effet de
frontière disparaisse – intégré à
indexmethods

03



03 - La construction de l'indice



CONTEXTE DE LA MISE EN PLACE DU MODÈLE

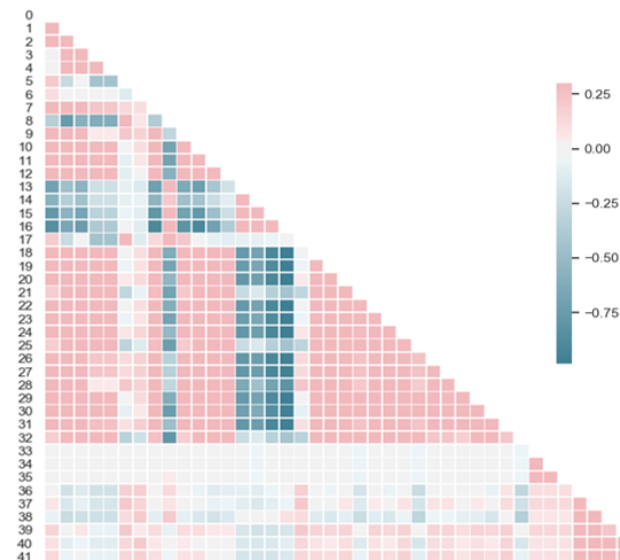
- 42 variables explicatives quantitatives
- Une variable réponse Y de nature inconnue correspondant au support du risque
- Exemples de supports du risque : production agricole / hectare, production solaire / MW.h⁻¹ installé, chiffre d'affaire d'une entreprise



INSUFFISANCE DE L'APPROCHE CLASSIQUE

- Premier reflexe : régression linéaire multiples, GLM
- Fortes hypothèses :
 - Liaisons linéaires à une transformation déterministe près
 - Indépendance entre les variables explicatives
 - Distribution de la réponse

Hypothèses non-vérifiées, et modélisation dépendante du support du risque





ALTERNATIVE : MÉTHODES PÉNALISANTES

- Lasso, Ridge, Elastic Net
- Sélectionne des modèles parcimonieux en terme de nombre de variables explicatives retenues
- « Gère » les corrélations en excluant, parmi un groupe de n variables corrélées, $n-1$ variables

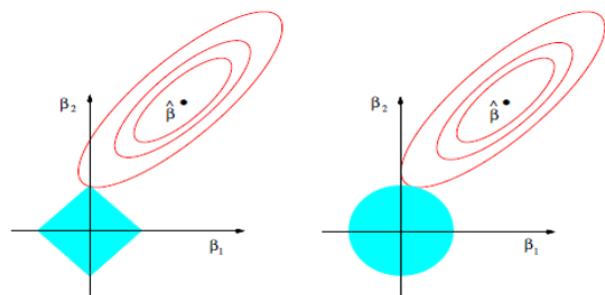


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.



ARBRES DE RÉGRESSION, FORÊTS ALÉATOIRES

- Modélisation de phénomènes non-linéaires
- Hyperparamétrage simple
- Le sur-apprentissage doit être contrôlé avec attention
- Ne prend en compte que des données tabulaires : les relevés de données distribués dans le temps et dans l'espace doivent être considérés comme des points de relevé indépendants
- Certains réseaux de neurones sont à même de modéliser la forte autocorrélation spatiale et temporelle des phénomènes météorologiques



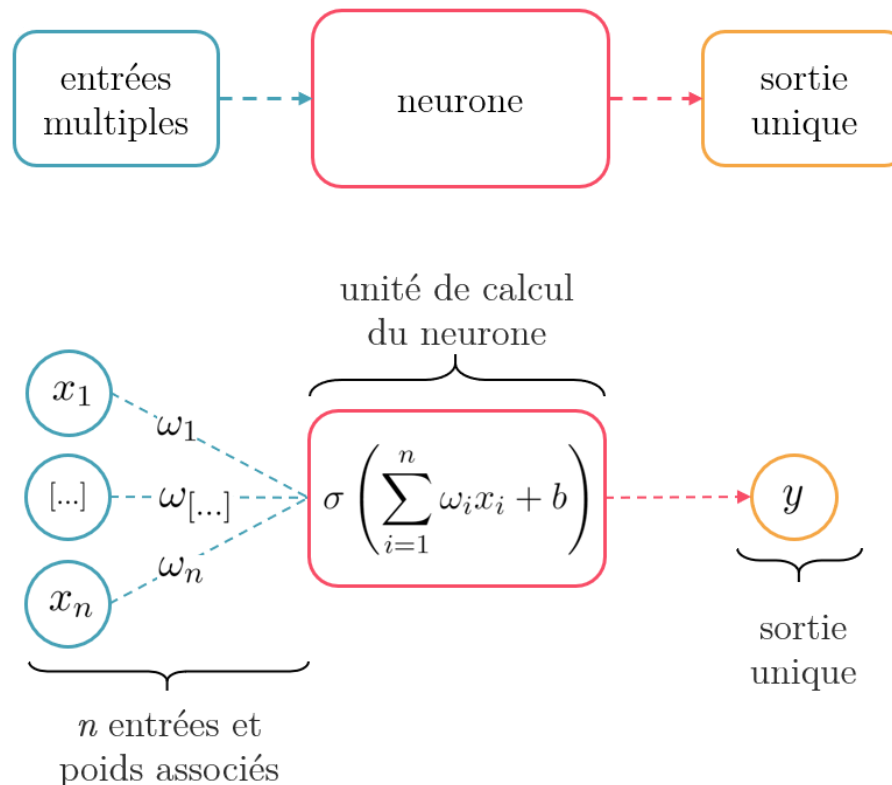
LE NEURONE ARTIFICIEL : PRINCIPE

- Le neurone fait une synthèse de l'information des multiples entrées pour obtenir une unique valeur de la sortie
- Il est possible que le neurone ne soit pas assez stimulé par les entrées pour être activé



LE NEURONE ARTIFICIEL : FORMELLEMENT

- Fonction de plusieurs variables retournant un unique scalaire
- Les poids servant à pondérer la somme et le terme de biais constituent les paramètres du neurones



Définition d'une couche de neurones artificiels



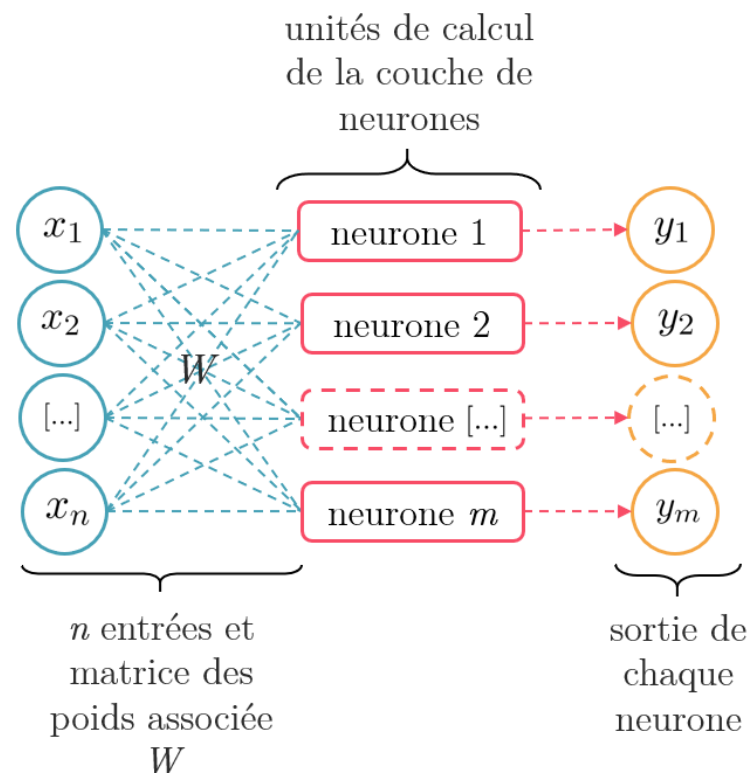
COMBINAISON DE NEURONES ARTIFICIELS : OBJECTIF

- Les neurones artificiels sont des unités simples (prédicteur linéaire auquel est appliquée une fonction)
- Ils présentent un intérêt lorsqu'ils sont combinés pour former des systèmes plus complexes

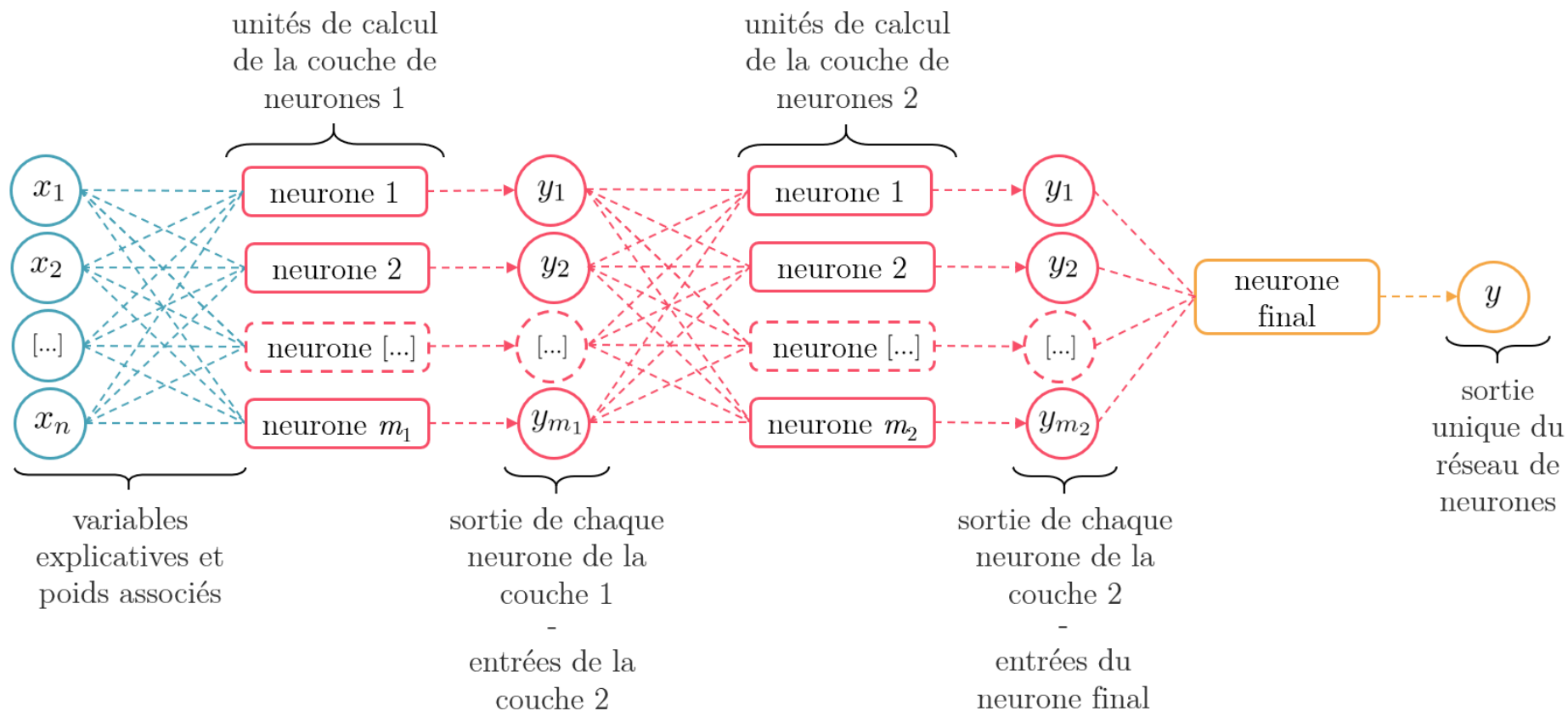


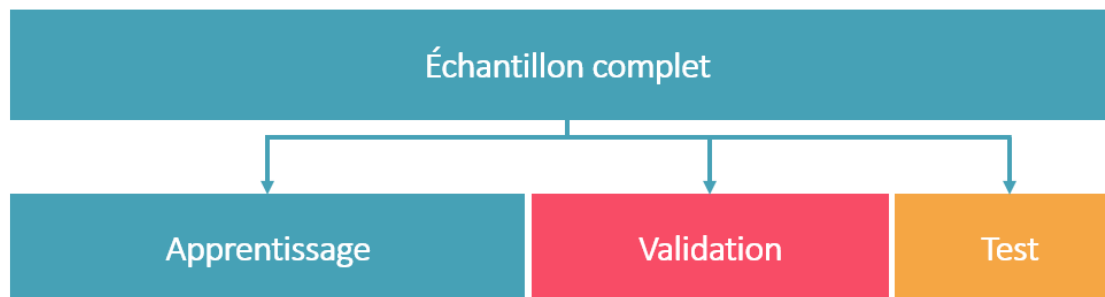
COUCHE DE NEURONES ARTIFICIELS : DÉFINITION

- Couche de m neurones : combinaison verticale de m neurones indépendants partageant les mêmes entrées
- Chaque neurone attribue des poids différents aux entrées et possède un biais qui lui est propre



Réseaux de neurones artificiels de type MLP





ERREUR D'APPRENTISSAGE : DÉFINITION

- Apprendre : capter un lien valable sur l'échantillon d'apprentissage entre les variables explicatives et la réponse
- Erreur d'apprentissage : MSE calculée sur l'échantillon d'apprentissage

$$MSE(\theta, h) = \frac{1}{p} \sum_{j=1}^p \left[y_j^{(pred; \theta; h)} - y_j \right]^2$$



ERREUR DE GÉNÉRALISATION : DÉFINITION

- Généraliser: le lien établi entre les variables explicatives et la réponse reste juste dans un contexte non-observé
- Erreur de généralisation : MSE calculée sur des l'échantillon de validation avec $\theta = \theta^*$

$$MSE(\theta^*, h) = \frac{1}{p} \sum_{j=1}^p \left[y_j^{(pred; \theta^*; h)} - y_j \right]^2$$



PARAMÈTRES : DÉFINITION

- Paramètres : poids et biais du réseau . ils influencent grandement la valeur de la sortie du réseau
- L'ensemble des paramètres est placé au sein d'un vecteur noté θ



PARAMÈTRES OPTIMAUX : DÉFINITION

- Définis comme les paramètres minimisant l'erreur d'apprentissage :

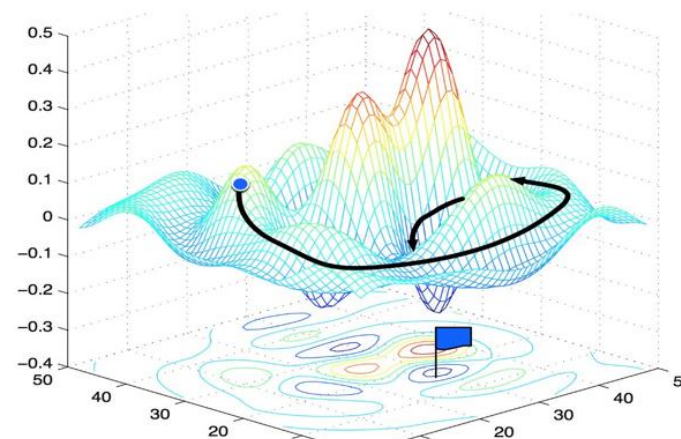
$$\theta^* = \underset{\theta}{\operatorname{argmin}} MSE(\theta)$$

- Déterminés par un algorithme d'optimisation itératif : ADAM



ADAM, SOMMAIREMENT

- Algorithme de descente de gradient populaire pour l'ajustement de réseaux de neurones
- Convergence rapide vers les paramètres optimaux



Descente de gradient vers le minimum global d'une fonction à deux variables



HYPERPARAMÈTRES : DÉFINITION

- Hyperparamètres : éléments devant être fixés avant la détermination des paramètres
- Sont liés à ADAM ou à l'architecture du réseau
- L'hyperparamétrage d'ADAM est automatisé : recherche itérative des hyperparamètres optimaux



ARCHITECTURE OPTIMALE D'UN RÉSEAU : DÉFINITION

- Architecture : agencement des neurones h
- Réseau MLP : nombre de couches et nombre de neurones par couche à déterminer
- Architecture optimale définie par :

$$h^* = \underset{h}{\operatorname{argmin}} MSE(h, \theta^*)$$



DÉTERMINATION DE L'ARCHITECTURE OPTIMALE : PROBLÉMATIQUE

- $MSE(h, \theta^*)$ est très coûteuse à évaluer
- Optimisation de $MSE(h, \theta^*)$ dans un espace de grande dimension : une recherche exhaustive ou aléatoire de h^* n'est pas envisageable (temps de calcul en semaines / mois sur un serveur)

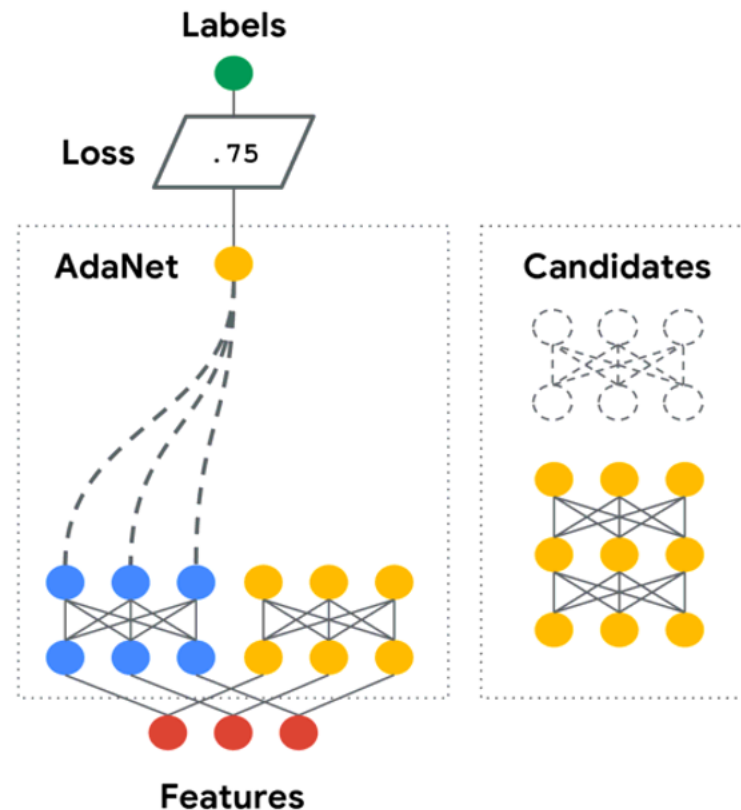


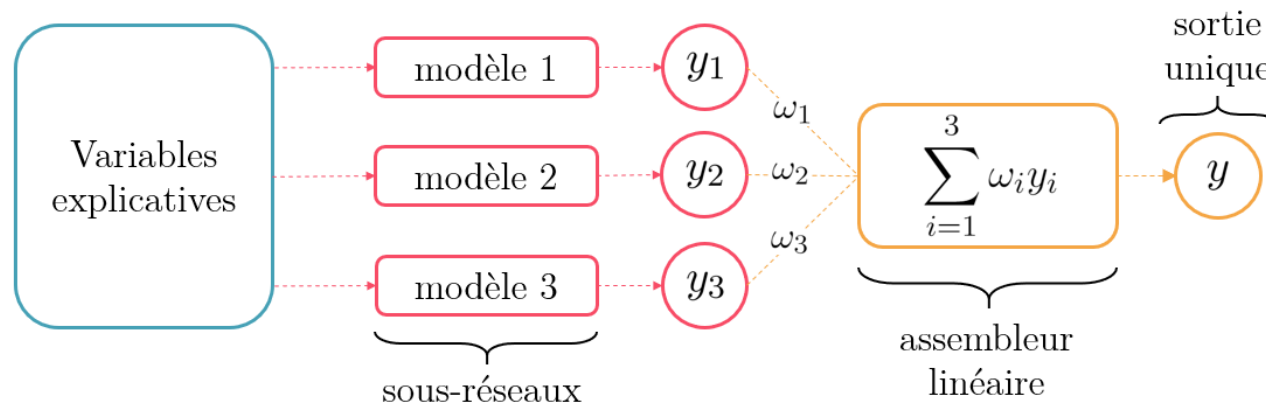
ADANET : ADAPTIVE STRUCTURAL LEARNING OF ARTIFICIAL NEURAL NETWORKS

- Article publié par Google Research à l'occasion de l'ICLM 2017
- Alternative à Auto ML (solution propriétaire et payante de Google)
- Implémentation sommaire disponible en Python ; lourde et peu flexible : son utilisation demande actuellement une expertise importante

SOLUTION OPÉRATIONNELLE : NOUVELLE IMPLÉMENTATION D'ADANET

- Entièrement automatisée
- Disponible au sein du package Python accompagnant ce mémoire : **indexmethods**
- Permet de mettre en place un réseau de neurones performant sans connaissances préalables du sujet





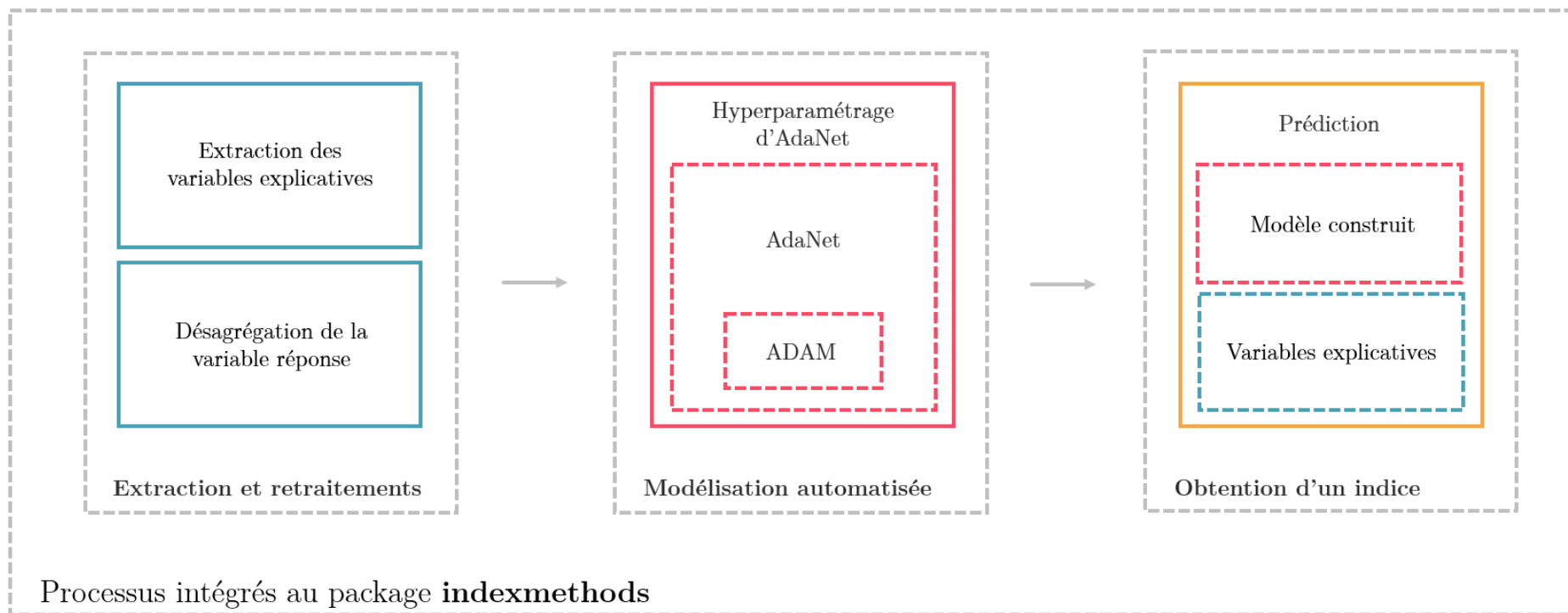
ENSEMBLE DE MODÈLES : DÉFINITION

- La sortie des modèles constituant l'ensemble est transformée en une sortie unique par un assembleur
- L'assembleur peut prendre des formes diverses



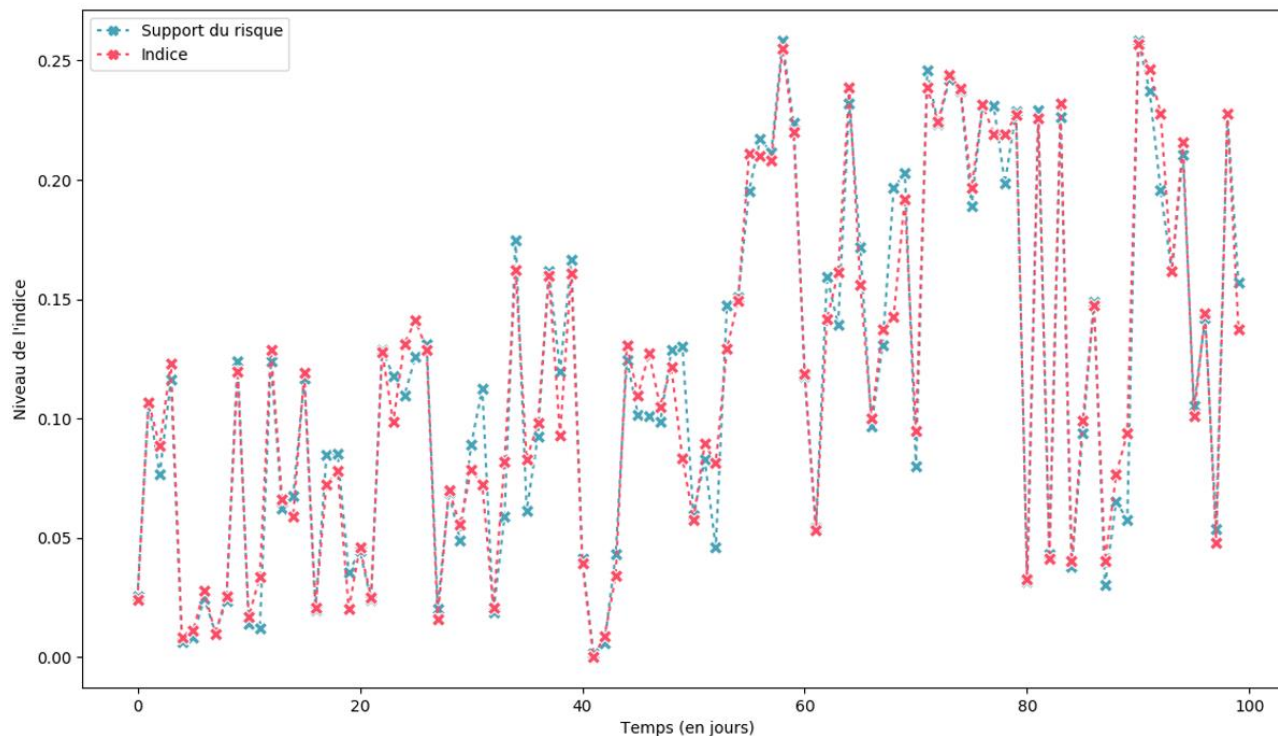
ADANET : CONSTRUCTION ITÉRATIVE D'ENSEMBLES

- Construit itérativement un ensemble de sous-réseaux
- Complexité croissante des sous-réseaux proposés pour ajout à l'ensemble
- Garantie théorique de convergence vers le maximum de généralisation



04 - Application à la création d'une couverture indiciaire

Indice pour la couverture des pertes d'exploitation des centrales solaires



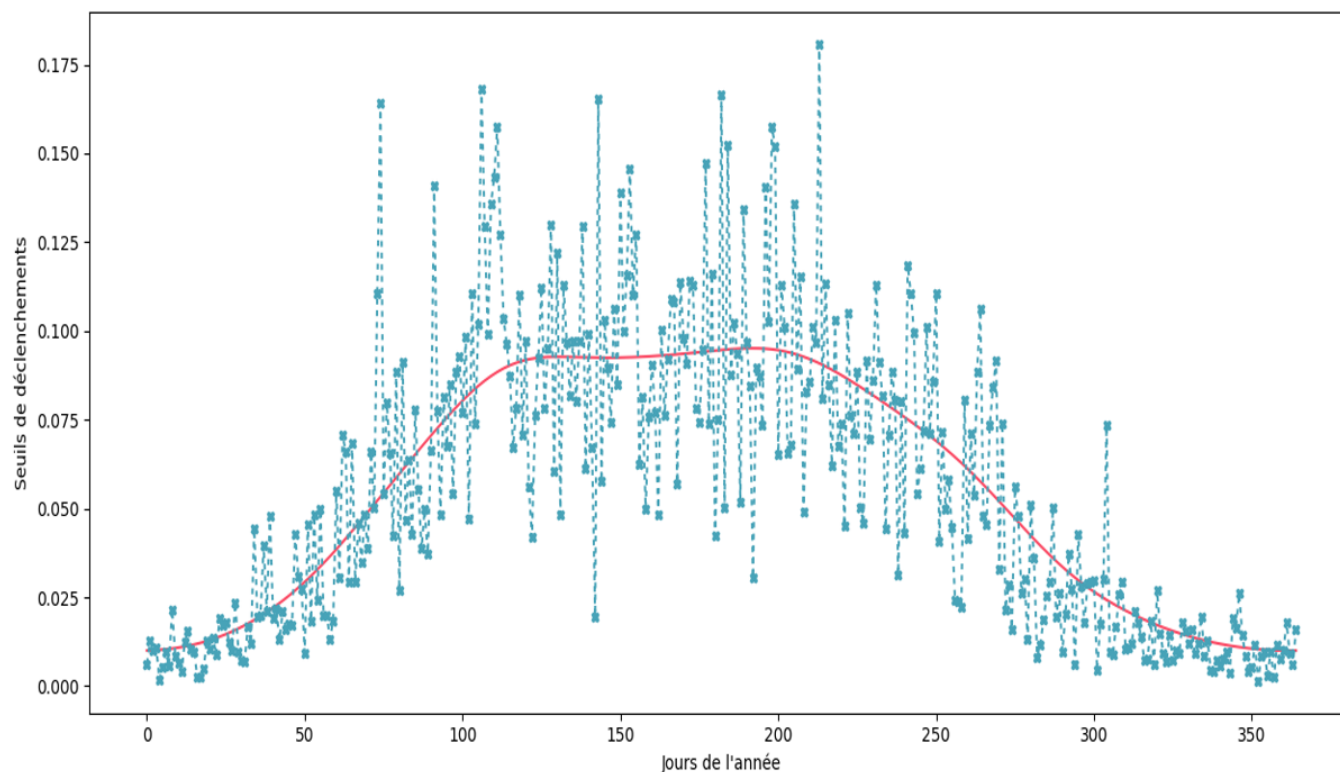
LE SUPPORT DU RISQUE

- Potentiel de production si 1MW/h de capacité était installée
- Sans unité

L'INDICE

- Prédiction du support du risque effectuée par AdaNet après sa convergence (3 jours)
- A été construit automatiquement à partir des variables météorologiques du programme POWER

Détermination des seuils de déclenchement

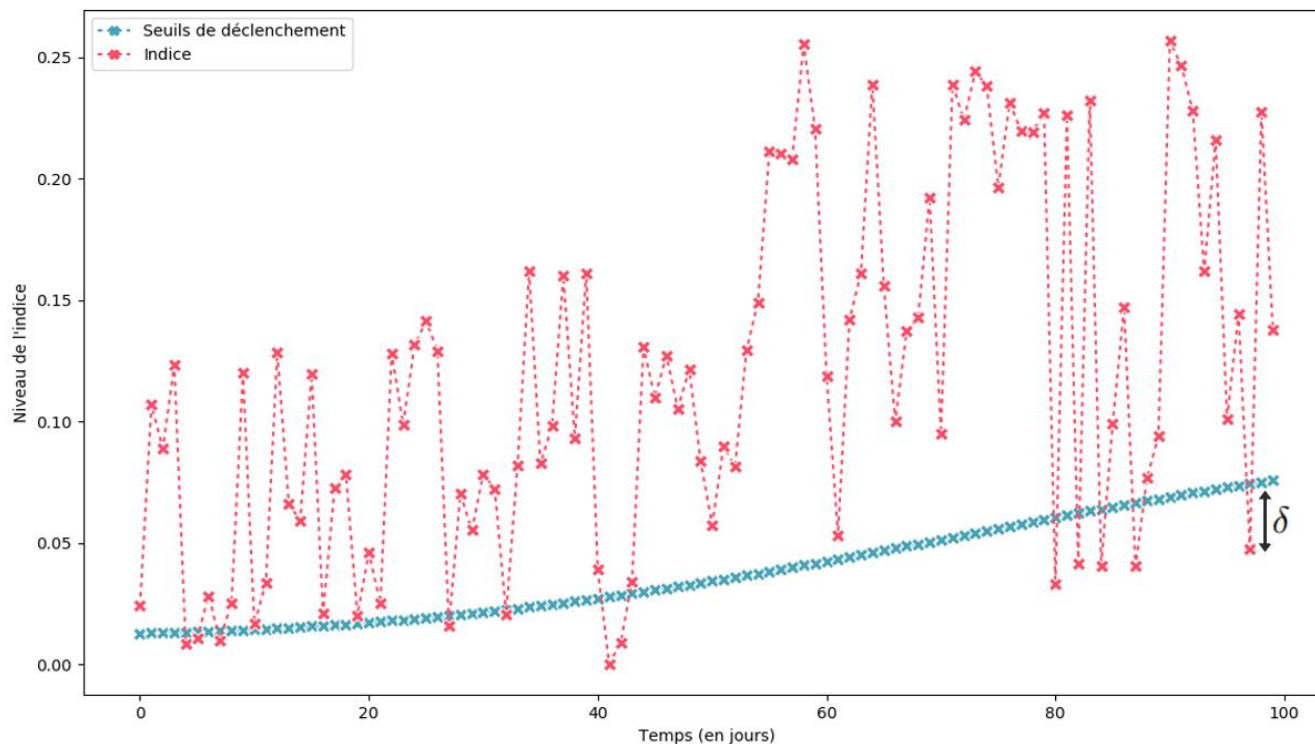


LES QUANTILES DE L'INDICE HISTORIQUE

- Le niveau du quantile reflète de niveau de couverture
- Un quantile est établi pour chaque jour de l'année

LISSAGE PAR MOYENNES MOBILES

- Les quantiles sont instables dans le temps (11 observations par jour)
- Une moyenne mobile uniforme symétrique est appliquée avec une fenêtre de 21 jours



SURVENANCE DU SINISTRE

- Définie comme le franchissement du seuil par l'indice
- Le nombre de sinistre moyen annuel peut-être déduit des seuils de déclenchement et de l'indice historique

MONTANT DE L'INDEMNISATION

- Peut être déduit du niveau de l'indice lorsqu'il dépasse le seuil

$$L = \delta \times \text{nombre de MW/h} \times \text{prix du MW/h}$$

05 - Et après ?

Extraction des
variables explicatives

Désagrégation de la
variable réponse

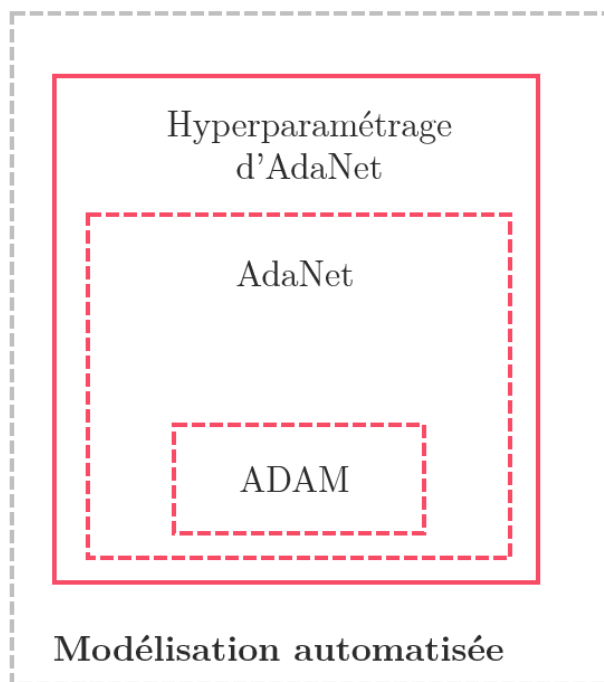
Extraction et retraitements

LA DÉSAGRÉGATION, UN PROBLÈME RÉCURRENT

- L'open data propose des informations à différentes mailles
- Nécessité de travailler la donnée pour obtenir une information uniforme
- La désagrégation est essentielle pour toute étude géospatiale



esa



AUTOMATISATION DE LA MISE EN PLACE DE RÉSEAUX MLP

- L'implémentation d'AdaNet effectuée au sein du package **indexmethods** peut être utilisée dans des contextes actuariels variés
- Ne requiert qu'un jeu de donnée contenant des variables explicatives quelconques et une variable réponse

→ TARIFICATION SANTÉ / AUTO / MRH

→ MODÈLES PRÉDICTIFS DANS DES
CONTEXTES DIVERS :
CLASSIFICATION & RÉGRESSION

Merci pour votre attention