Identification d'une structure de réseau au service de l'anticipation des risques
Exemples en catastrophes naturelles et en risque cyber

Geoffrey Ecoto

Responsable Actuariat & Provisionnement - CCR

Olivier Lopez

Professeur des Universités

Sorbonne Université & Detralytics

Maud Thomas

Maitre de conférences

Sorbonne Université

La théorie classique du risque

- Considérons un portefeuille de n assurés avec $(Y_i)_{1\leq i\leq n}$ leurs pertes associées.
- Supposons que les $(Y_i)_{1\leq i\leq n}$ sont **indépendants** et identiquement distribués.
- La perte totale du portefeuille est

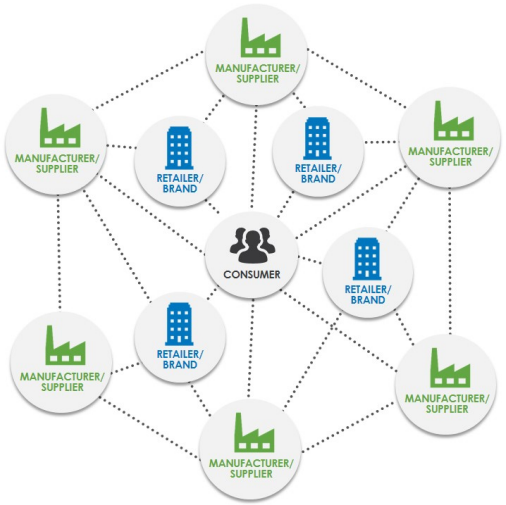
$$L = \sum_{i=1}^{n} Y_i$$

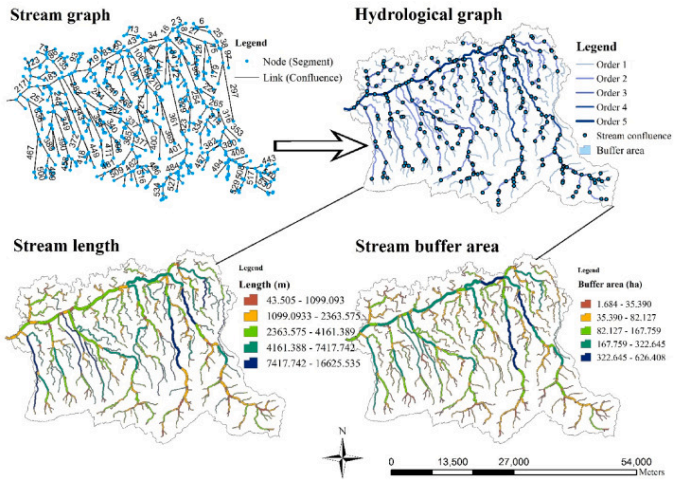
- Si $E[Y_1^2] < \infty,$ alors d'après le théorème limite central,

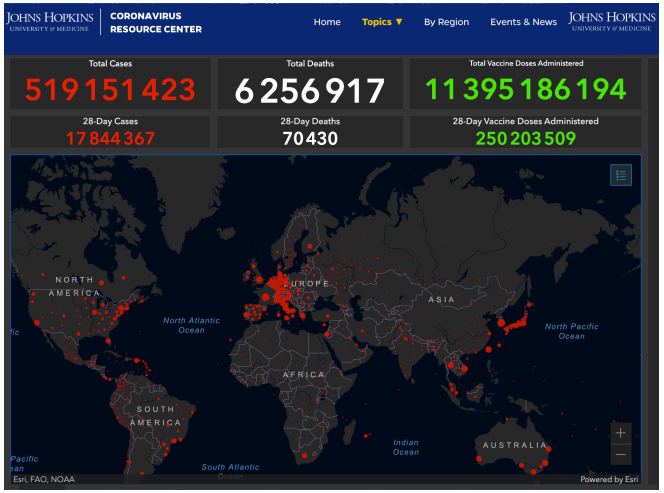
$$L \approx \mathcal{N}\left(n\pi, n\sigma^2\right)$$

où $\pi = E[Y_1]$ et $\sigma^2 = Var(Y_1)$


Digital Attack Map
Top daily DDoS attacks worldwide
Map · Gallery · Understanding DDoS · FAQ · About
United States
EC
Brazil
TH


MANUFACTURER/ SUPPLIER
RETAILER/ BRAND
CONSUMER


Stream graph
Legend
Node (Segment)
Link (Confluence)
Hydrological graph
Legend
Order 1
Order 2
Order 3
Order 4
Order 5
Stream confluence
Buffer area
Stream length
Length (m)
43.505 - 1099.093
1099.0933 - 2363.575
2363.575 - 4161.389
4161.388 - 7417.742
7417.742 - 16625.535
Stream buffer area
Buffer area (ha)
1.684 - 35.390
35.390 - 82.127
82.127 - 167.759
167.759 - 322.645
322.645 - 626.408
N
0
13,500
27,000
54,000
Meters


JOHNS HOPKINS UNIVERSITY & MEDICINE
CORONAVIRUS RESOURCE CENTER
Home
Topics
By Region
Events & News
Total Cases
519 151 423
Total Deaths
6 256 917
Total Vaccine Doses Administered
11 395 186 194
28-Day Cases
17 844 367
28-Day Deaths
70 430
28-Day Vaccine Doses Administered
250 203 509
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
North Atlantic Ocean
North Pacific Ocean
South Atlantic Ocean
Indian Ocean
Esri, FAO, NOAA
Powered by Esri

RESEAU-NENS

Plan

Partie I - Risques naturels

1. Le régime d'indemnisation de catastrophes naturelles : le cas de la sécheresse
2. L'étude d'un algorithme de super learning : les graphes pour la modélisation des dépendances spatiales

Partie II - Le risque cyber

1. Lien avec les phénomènes de cotation
2. Modèle à variable cachée
3. Identification d'une structure de graphe implicite

Part I
Risques naturels

1. Le régime d'indemnisation des catastrophes naturelles et le risque sécheresse
2. L'étude d'un algorithme de super learning : les graphes pour la modélisation des dépendances spatiales

Risques naturels

Le régime d'indemnisation des catastrophes naturelles et le risque sécheresse

Le régime d'indemnisation des catastrophes naturelles

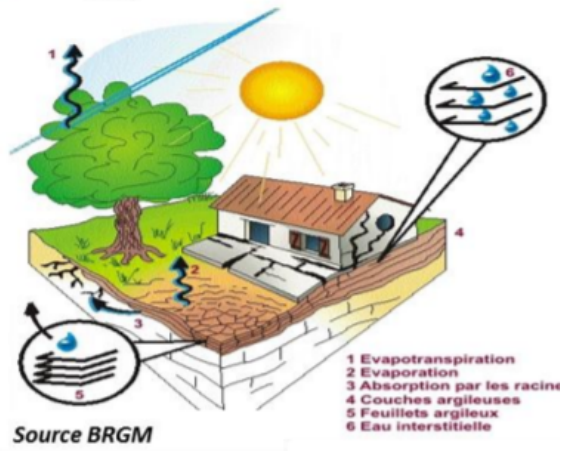
- Une extension de garantie obligatoire des contrats DAB
- La Caisse Centrale de Réassurance (CCR) indemnise les assureurs dans le cadre de ce régime si les biens assurés sont situés dans une commune ayant fait l'objet d'une reconnaissance de l'état de catastrophe naturelle de la part de la commission interministérielle
- Les maires formulent les demandes de reconnaissance

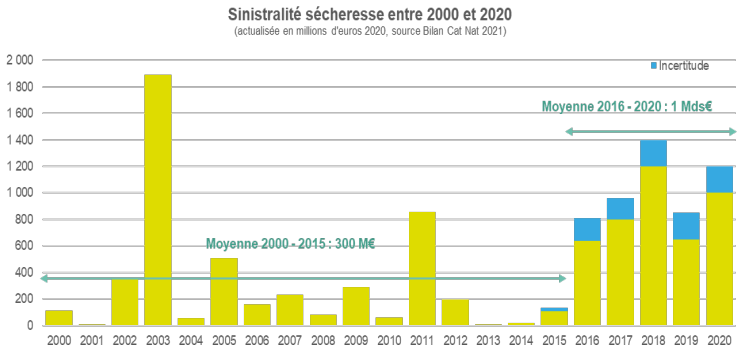
La sécheresse géotechnique : définition et enjeux financiers

- Retrait et gonflement des argiles (RGA) dans le sol provoquant des fissures sur les maisons
- Second péril le plus couteux après les inondations, avec une sinistralité annuelle moyenne de 475 M€

Dans le cadre de ses travaux, CCR est amenée à estimer le coût d'un événement sécheresse

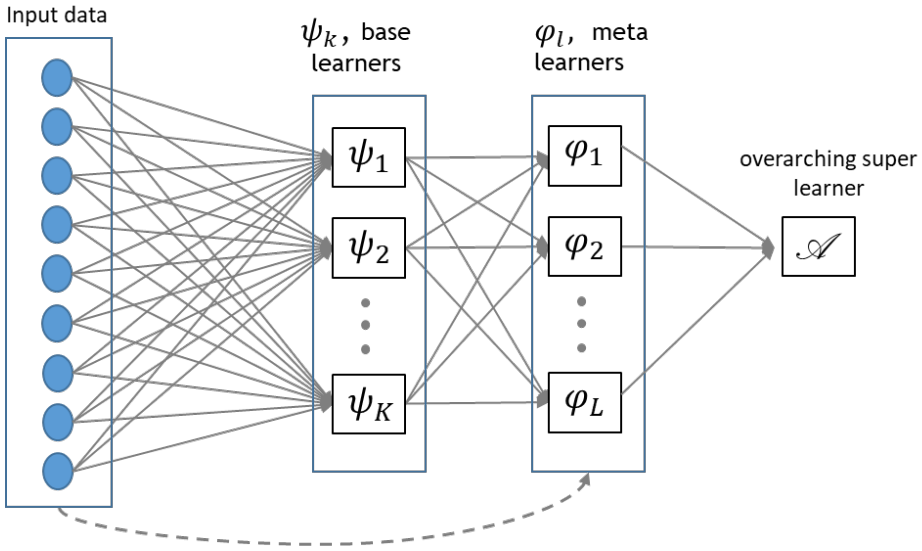
- Un modèle en production (régressions logistiques)
- Volonté de CCR de s'approprier les méthodes de machine learning via l'initiation d'une thèse en partenariat avec le MAP5 encadrée par Antoine Chambaz (MAP5) et Thierry Cohignac (CCR). Soutenance prévue début 2023.

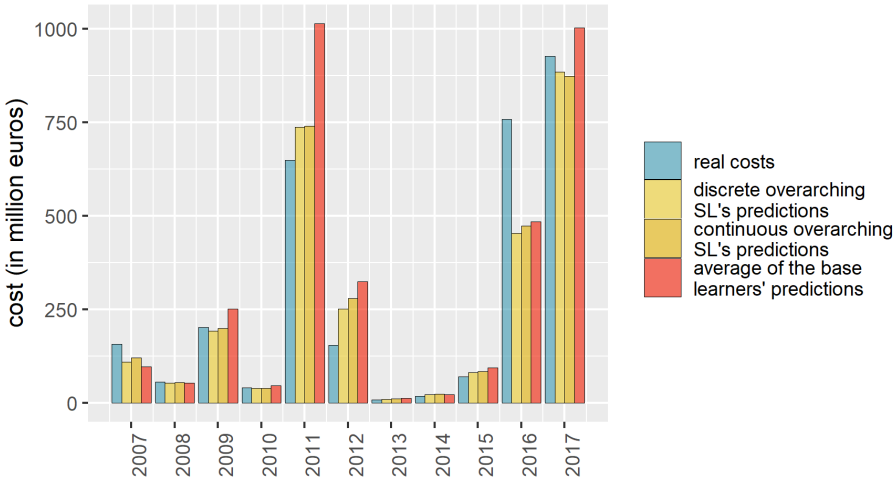




Risques naturels

Développement d'une nouvelle méthodologie pour l'estimation des dommages provoqués par la sécheresse géotechnique





Coûts réels des événements sécheresse 2007 à 2017, prédictions réalisées par l'overarching super learner, moyennes des prédictions des base learners

- Une méthodologie inspirée du Super Learner (van der Laan et al. 2007), un algorithme d'agrégation de modèles possédant des garanties théoriques renseignant sa capacité à sélectionner ou combiner le ou les bons algorithmes fondamentaux (base learners)
- Ces résultats théoriques reposent sur une hypothèse d'indépendance des observations
- Dans le cadre de notre étude, comment gérer la dépendance spatiale entre les observations ?

Risques naturels

La sécheresse géotechnique et les dépendances spatiales

- Le Super Learner réalise sa sélection parmi les algorithmes fondamentaux en se basant sur le risque empirique
- Grâce à l'utilisation d'une inégalité de concentration de Janson (2004) modélisant la structure de dépendance entre les observations grâce à un réseau, nous obtenons le résultat (simplifié) suivant : pour tout $\varepsilon > 0$,

$$Risque\left(\hat{\psi}\right) - (1+\varepsilon)\ .\ Risque\left(\tilde{\psi}\right) \leq constante(\varepsilon)\ .\left(\frac{log\left(K\ .\ log\left(\mathcal{I}^2\right)\right)}{\mathcal{I}^2}\right)$$

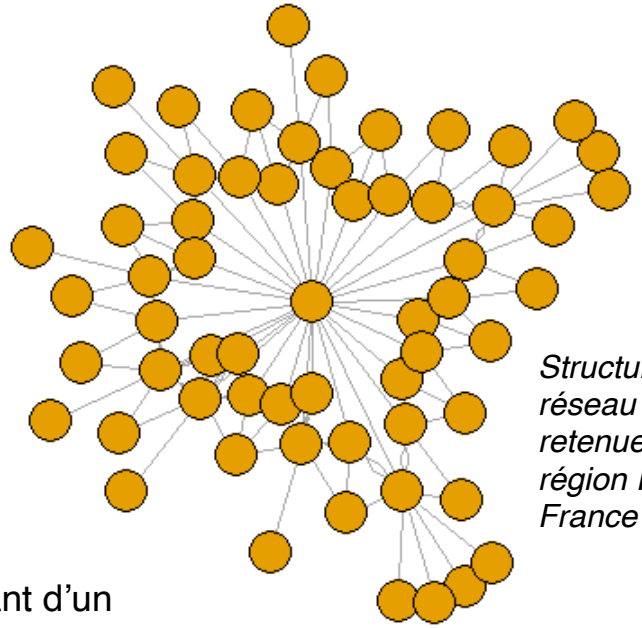
- $\hat{\psi}$ et $\tilde{\psi}$ sont respectivement l'algorithme sélectionné par le Super Learner et l'algorithme oracle (optimal)
- K est le nombre d'algorithmes fondamentaux
- $\mathcal{I}$ est une quantité d'information. Naturellement, nous souhaitons que $\mathcal{I}$ soit la plus grande possible

➔ $\mathcal{I}$ dépend d'une caractéristique du réseau…

Risques naturels

La sécheresse géotechnique et les dépendances spatiales

- Soit $\mathscr{A}$ l'ensemble des communes de France métropolitaine ($\left|\mathscr{A}\right|$ est de l'ordre de 36 000)
- Dans le cadre de notre étude, nous introduisons le réseau des dépendances des communes françaises $\mathscr{G}$ incorporant la mitoyenneté entre communes

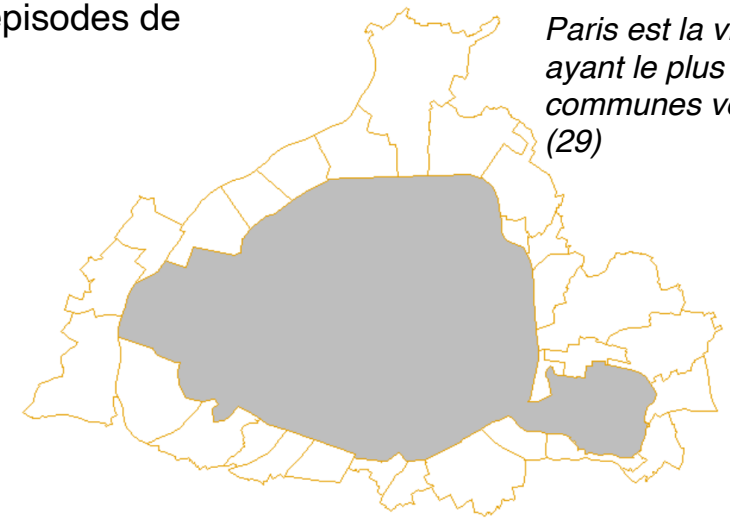


Structure de réseau retenue – région Ile-de-France

- Le degré $deg(\mathscr{G})$ de $\mathscr{G}$ (1 plus le nombre maximum d'arêtes partant d'un nœud de $\mathscr{G}$) quantifie la prégnance de la dépendance
- $\mathscr{I} = \left|\mathscr{A}\right| / (t \,.\, deg(\mathscr{G}))$, où $t > 1$ désigne le nombre d'épisodes de sécheresse observés

min.	1st qu.	median	mean	3rd qu.	99%-qu.	max
0	5	6	5.96	7	11	29

Quartiles, quantile 99% et moyenne du nombre de communes voisines en France en 2019

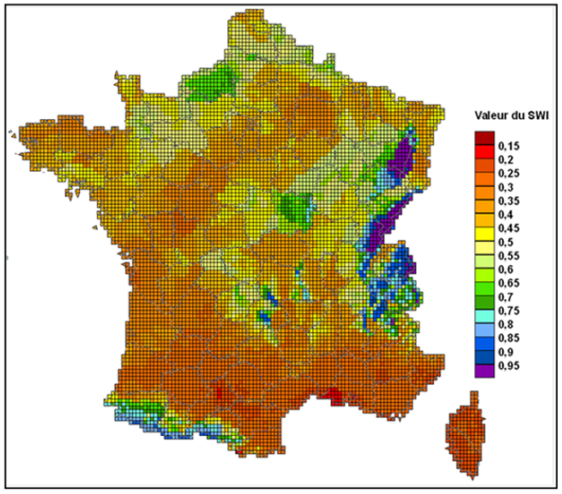


Paris est la ville ayant le plus de communes voisines (29)

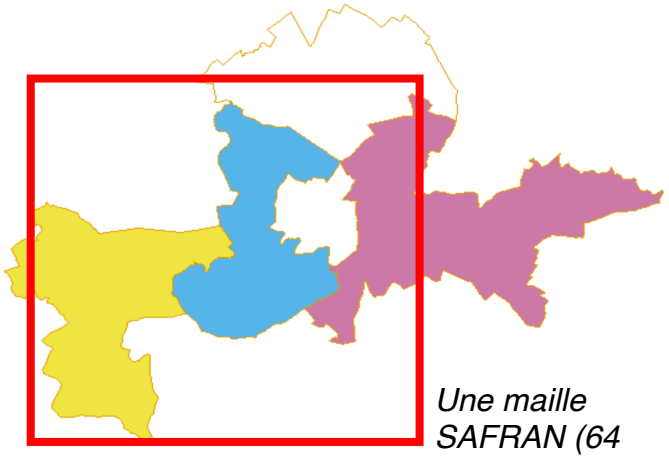
Risques naturels

La sécheresse géotechnique et les dépendances spatiales

- Ces justifications, théoriques (pas de rôle opérationnel), indiquent qu'il est possible de réaliser l'apprentissage souhaité malgré une série chronologique courte (une trentaine d'épisodes de sécheresse observés) composée d'un grand nombre d'observations spatialement dépendantes à chaque pas de temps
- Il existe des limites au recours à cette structure de réseau basée sur la mitoyenneté :
  - Les conditions météorologiques liées à l'humidité du sol sont représentées par le Soil Wetness Index (SWI)
  - La résolution de cet indice est de 64 km² (mailles SAFRAN) quand la superficie moyenne des communes françaises est de 14 km² : la mitoyenneté ne peut modéliser la dépendance spatiale engendrée par la nature météorologique du phénomène étudié
  - Si la structure de réseau permet notamment de modéliser les dépendances liées aux aspects comportementaux, nous avons recours au conditionnement pour le prise en compte des dépendances spatiales liées aux conditions météorologiques



Indice d'humidité du sol


Une maille SAFRAN (64 km²)

Part II
Risque cyber

1. Lien avec les phénomènes de cotation
2. Modèle à variable cachée
3. Identification d'une structure de graphe implicite

Un exemple : Wannacry

- Rançongiciel Wannacry : cyberattaque mondiale en mai 2017.
- Elle a utilisé la vulnérabilité "EternalBlue".
- Environ 200 000 ordinateurs infectés à travers 150 pays sur environ une semaine.
- Estimation du coût : des centaines de millions de dollars, des milliards selon certaines estimations. (100 millions de livres sterling pour le NHS).

Autre exemple : Colonial Pipeline (2021)

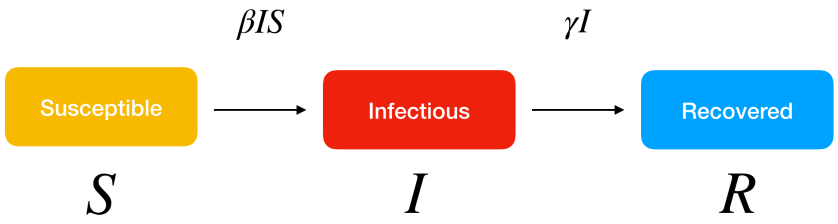
- Hausse de 4,2% du WTI et du Brent.
- "Double extorsion" : attaque par rançongiciel combinée à du chantage.
- Auteurs : le groupe de hackers "Darkside" (Rançongiciel sous forme de service).

Loi du temps d'infection

- T = date d'infection de l'assuré par l'attaque informatique
- Définition par le risque instantané

$$\lambda_T(t) = \lim_{t \to 0} \frac{\mathbb{P}(T \in [t, t+dt] \,|\, T \geq t)}{dt}.$$

- Nous pouvons introduire un modèle épidémiologique pour décrire la propagation de l'épidémie dans la population mondiale (pas seulement sur le portefeuille),
- Et, déduire $\lambda_T(t)$ de ce modèle



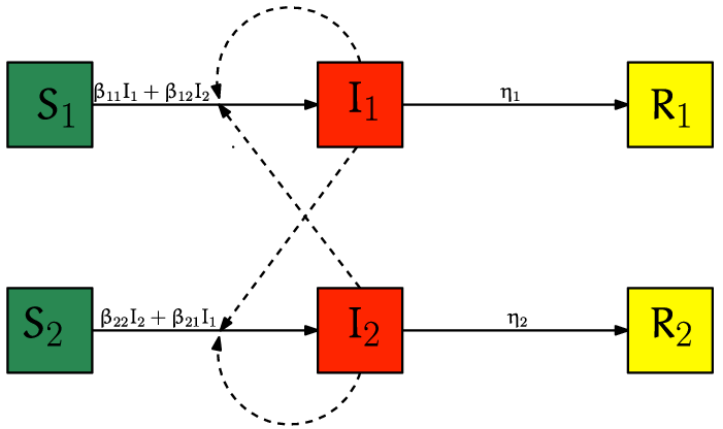
Modèle multi-SIR

- Les assurés sont regroupés en $j = 1,\ldots,d$ catégories.
- Système d'équations différentielles :

$$\frac{ds_j(t)}{dt} = -\left\{\alpha_j(t) - \sum_{k=1}^{d} \beta_{k,j} i_k(t)\right\} s_j(t),$$

$$\frac{di_j(t)}{dt} = \left\{\alpha_j(t) + \sum_{k=1}^{d} \beta_{k,j} i_k(t)\right\} s_j(t) - \gamma_j i_j(t)$$

$$\frac{dr_j(t)}{dt} = \gamma_j i_j(t)$$



Comment calibrer la matrice de contagion ?

- Dans le document (voir les références à la fin), nous utilisons la matrice suivante :

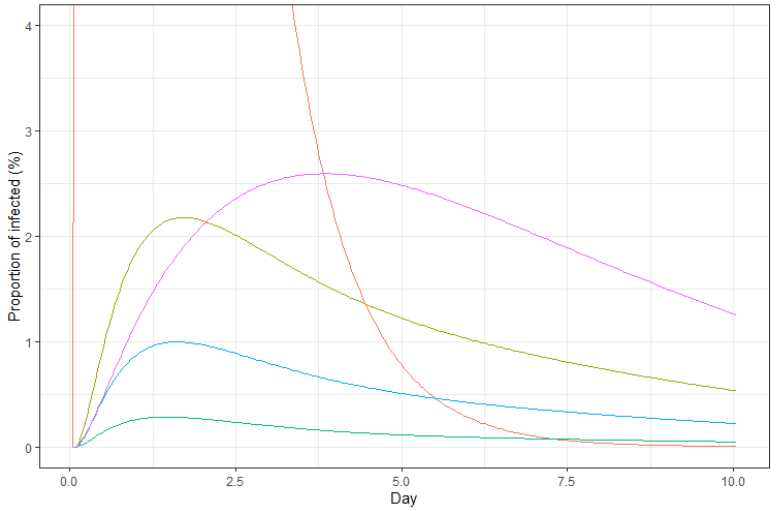
$$B = (\beta_{k,j})_{1 \leq k,j \leq d}$$

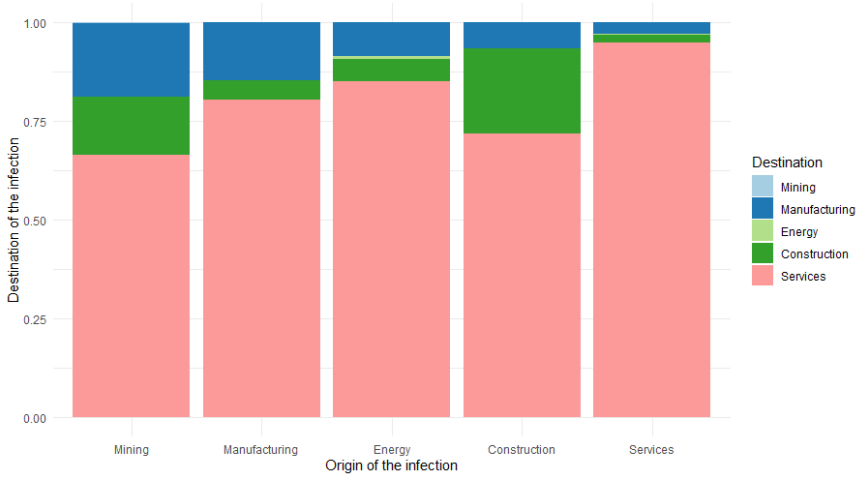
	Mining	Manufacturing	Energy	Construction	Services	Total
Mining	0,0634	0,2927	0,0449	0,1427	0,1255	0,6692
Manufacturing	0,0063	0,0527	0,0027	0,0108	0,0351	0,1076
Energy	0,0135	0,0370	0,0571	0,0150	0,0452	0,1679
Construction	0,0019	0,0068	0,0007	0,0141	0,0091	0,0326
Services	0,0003	0,0042	0,0004	0,0017	0,0161	0,0227
Total	0,0855	0,3934	0,1057	0,1844	0,2309	1

- Avertissement : cette matrice n'a qu'un but d'illustration.
- Une méthode de calibration est proposée, basée sur une petite quantité de données macroscopiques pour matérialiser l'intensité des connexions entre les secteurs

Quelques résultats typiques

- On peut étudier l'effet de différents types d'attaques.
- Ici, simulation d'un épisode Wannacry ciblant un secteur économique particulier.





Estimation d'un graphe

- L'estimation de la matrice B peut être identifiée via l'estimation d'un graphe valué.
- Exemple de stratégie : inspiré des méthodologies utilisées en écologie, voir par exemple Chiquet, Mariadassou, Robin (2021).
- Dans cet article, les auteurs étudient l'abondance des espèces dans différentes zones.
- Analogie : lorsque la catégorie k est fortement touchée par les cyberattaques, est-ce également le cas pour la catégorie j ?
- Lien avec l'analyse de corrélation mais adapté aux données de comptage

Modèle Poisson log-normal (PLN)

- Soit $N_{i,j}$ le nombre de sinistres de la catégorie j pour la période i.
- Introduisons un vecteur aléatoire caché $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,d})$
- Supposons que :
  - $N_{i,j} \,|\, \mathbf{Z}_i = \mathbf{z}$ est distribuée selon une loi de Poisson avec une moyenne de $\exp(\mu_j + z_j)$
  - $\mathbf{Z}_i \sim \mathcal{N}(0, f(B))$
- Reste à estimer B .
- **Question** : Quel est le lien entre ce modèle et le modèle épidémiologique précédent ?

Lien entre les deux modèles

- Considérons le modèle épidémiologique précédent
- Hypothèse 1 : la matrice de connexion B est **symétrique**
- Hypothèse 2 : les connexions implicites entre catégories d'assurés sont les mêmes en situation de "bas régime " qu'en phase épidémique
- Ensuite, on suppose que le nombre de sinistres $N_{i,j}$ est produit par :
  - des attaques qui frappent indépendamment chaque secteur
  - de petites cyber-épidémies utilisant le même schéma de contagion B
- Alors, le modèle PLN peut être vu comme une approximation du modèle contagieux précédent, et on peut retrouver B, en utilisant l'inférence dans le cadre PLN.

Inférence dans le modèle PLN

- Données incomplètes

  —> l'idée serait d'utiliser l'**algorithme EM.**

- L'étape E nécessite le calcul de la loi de $\mathbf{Z}_i \,|\, N_{i,j}$, qui n'a pas de formule fermée.
- Alternative : utiliser l'approximation variationnelle, en se liant à un algorithme VEM (variational EM).
- Le paquetage R PLNmodels est disponible pour effectuer cette tâche.
- Comprend une régularisation (LASSO) pour identifier les structures de graphe parcimonieux

Références

- Detra Note: *Modeling accumulation scenarios in cyber risk*, see https://detralytics.com/detra-notes/ (with L. d'Oultremont and B. Sporrenberg)

- Detralytics tool (preview): https://detralytics.shinyapps.io/cyber_risk/

- Hillairet, C., Lopez, O. (2021) *Propagation of cyber incidents in an insurance portfolio: counting processes combined with compartmental epidemiological models,* **Scand. Actuarial Journal.** https://www.tandfonline.com/doi/abs/10.1080/03461238.2021.1872694

- Hillairet, C., Lopez, O., d'Oultremont, L., Spoorenberg, B. (2022) *Cyber contagion: impact of the network structure on the losses of an insurance portfolio,* Preprint, https://hal.archives-ouvertes.fr/hal-03388840/

- Chiquet, J., Mariadassou, M., & Robin, S. (2021). *The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances,* **Frontiers in Ecology and Evolution, 9, 588292.**