INSPQ INSTITUT NATIONAL DE SANTÉ PUBLIQUE DU QUÉBEC

Modélisation de pics sanitaires à l'aide de la théorie des valeurs extrêmes

BiESP BUREAU D'INFORMATION ET D'ÉTUDES EN SANTÉ DES POPULATIONS

SURVEILLANCE DES MALADIES CHRONIQUES Numéro 23

PRINCIPAUX CONSTATS

En santé publique, les analyses des décès et des hospitalisations sont habituellement effectuées en modélisant l'ensemble des données. Cette pratique est toutefois inadaptée pour appréhender les pics sanitaires, qui constituent des évènements « inhabituels » avec un impact majeur sur le système de services de soins de santé.

Ce rapport présente la façon de modéliser les séries de pics sanitaires à l'aide de la théorie des valeurs extrêmes, peu utilisée en surveillance de la santé.

L'extraction des pics est effectuée par la méthode des blocs et celle du seuil élevé; ces pics étant modélisés après plusieurs étapes de prétraitement des données et divers tests statistiques.

Mise en contexte

Dans le domaine de la santé, l'analyse de données statutaire, dont les décès et les hospitalisations, caractérise généralement le comportement de l'ensemble des données d'une série, et en particulier celui qui est observé en moyenne.

Par exemple, « au Québec, de 1989 à 2006, le taux quotidien d'admissions hospitalières pour maladies ischémiques cardiaques chez les personnes âgées d'au moins 65 ans était plus élevé lors des premières journées de températures élevées d'une saison, surtout lorsque ces températures perduraient quelques jours » (Bayentin et collab., 2010). Cette façon d'analyser les données a prouvé maintes fois son utilité au fil du temps. Les résultats précédents, par exemple, permettent d'identifier certaines conditions météorologiques menant à plus d'hospitalisations pour cause cardiovasculaire chez les aînés, et de rappeler aux services publics du Québec l'importance d'émettre les messages de santé et sécurité lors de leur survenue. Ce faisant, l'importance des pics est diluée dans ce qui est observé en moyenne dans les données. De plus, les caractéristiques des pics, qui auraient pu les distinguer et permettre d'analyser les variables environnementales ou socio-économiques associées à leur survenue, sont plutôt assimilées à la tendance.

Or, l'étude du comportement des séries de pics, encore inexploitée dans le domaine de la santé, peut présenter un grand intérêt. Elle pourrait notamment soutenir les gestionnaires municipaux et de la santé lors de situations d'urgence (pics d'interventions psychosociales, d'hospitalisations, de consultations, etc.), qui nécessitent une planification rigoureuse afin de déployer des ressources humaines et matérielles rapidement, au bon endroit et au moment opportun. L'étude des séries de pics profiterait aussi à la surveillance de la santé des populations, notamment en circonscrivant davantage les éclosions de maladies infectieuses (influenza durant l'hiver,

Québec

gastroentérites lors de fortes pluies, etc.) et les pics de maladies chroniques (mortalité cardiovasculaire lors d'une canicule, crises d'asthme lors de froid intense, etc.). Enfin, elle permettrait de déterminer les groupes à risque qui leur sont associés, lesquels se distinguent peut-être de ceux déjà identifiés avec les tendances et puis, le cas échéant, de mieux cibler les interventions de santé publique.

L'objectif de ce projet est de suggérer une façon d'appréhender les pics sanitaires à l'aide de techniques statistiques appropriées dans une perspective de surveillance en santé publique.

Les pics étudiés touchent les maladies cardiovasculaires (MCV), qui constituent la deuxième cause de mortalité (Girard et collab., 2016), la première cause d'hospitalisations au Québec (Daigle, 2007), ainsi que la catégorie de maladies la plus coûteuse pour le système de santé. Par MCV, on entend les maladies coronariennes, les maladies vasculaires cérébrales et l'insuffisance cardiaque. Dans ce projet, l'issue sanitaire considérée est le nombre de jours d'hospitalisation pour cause principale de MCV dans les régions métropolitaines de recensement (RMR) de Montréal et de Québec, de 1996 à 2007. Les données sanitaires proviennent de l'Institut national de santé publique du Québec (INSPQ)[1].

L'ensemble de la démarche a été utilisé avec succès dans le Programme de recherche en santé cardiovasculaire et changements climatiques 2011-2016, réalisé dans le cadre d'une entente entre l'INSPQ et le Centre de recherche Eau Terre Environnement de l'Institut national de recherche scientifique du Québec. Ces étapes sont présentées de façon allégée dans ce rapport. Pour plus de détails, voir les documents sources (Chiu et collab., 2015; Chiu et collab., 2016).

Quelle théorie utiliser pour l'étude des pics sanitaires?

La théorie des valeurs extrêmes (ou *extreme value theory* en anglais) a été retenue à titre d'assise théorique pour l'ensemble de la démarche méthodologique. Cette théorie s'intéresse aux valeurs extrêmes des distributions de probabilité et s'appuie sur des bases bien établies (Coles, 2001; Reiss & Thomas, 2007)[2]. En outre, elle a fait l'objet de nombreuses applications dans les domaines de l'hydrologie, de l'assurance et de la finance.

Parmi les rares exemples, De Zea Bermudez & Mendes (2012) ont utilisé la théorie des valeurs extrêmes pour modéliser les niveaux élevés de cholestérol chez les Portugais avec un seuil de pic fixé au quantile à 90 %. (Watts, Dupuis & Jones 2006) ont appliqué cette théorie pour estimer la borne supérieure de l'espérance de vie sur les populations canadienne et japonaise. Guillou, Kratz & Le Strat (2014) y ont fait appel pour étudier la répartition temporelle de *Salmonella* en France à des fins de surveillance. Enfin, Chen et collab., (2015) l'ont utilisé pour prédire les niveaux de retour de la grippe en Chine. Selon tous ces auteurs, la théorie des valeurs extrêmes est adaptée à l'étude des pics sanitaires. De l'avis de certains, la notion de niveau de retour, qui dépend de la période de retour, s'avère aussi d'intérêt.

La « période de retour » indique le temps moyen que mettra une variable avant d'atteindre ou de dépasser un certain niveau sur une longue période. Elle permet donc d'apprécier le caractère de rareté d'un évènement inhabituel. Cette notion est répandue en hydrologie (Salvadori et collab.; 2011) et en finance (Longin, 2000). Par exemple, en hydrologie, une période de retour de 100 ans pour une crue signifie que cette crue (centennale) a une chance d'être observée en moyenne tous les 100 ans. Cette information est alors d'importance vitale pour l'élaboration des ouvrages hydrauliques, et vraisemblablement aussi pour d'autres types d'ouvrages, comme la conception d'établissements de santé tenant compte des changements climatiques.

---

1 Banque de fichiers médico-administratifs jumelés obtenus avec l'accord de la Commission d'accès à l'information du Québec (CAI) le 27 mars 2008 (# 11 09 45 (08 19 12, 07 00 76)) dans le cadre de l'Étude de la faisabilité de développer la surveillance des MCV au Québec (1996-2007).

2 À notre connaissance, l'ouvrage de Reiss & Thomas est le seul qui consacre un court chapitre (chapitre 19) à l'application de la théorie des valeurs extrêmes en santé publique.

De façon symétrique à la période de retour, il est possible d'interpréter le « niveau de retour ». Il s'agit du niveau atteint ou dépassé en moyenne une fois sur une période donnée et pouvant être vu comme un quantile. Dès lors, le niveau de retour peut s'appliquer à divers secteurs de la santé, par exemple en offrant une mesure concrète pour la gestion optimale de la capacité en lits dont devrait disposer un établissement de santé sur la prochaine année ou sur une plus longue période. Son utilisation en surveillance de la santé des populations, au moment de déterminer des seuils critiques pour l'intervention de protection de la santé publique, en est un autre exemple.

Le calcul du niveau de retour dépend de la distribution des pics sanitaires.

Comment extraire les séries de pics de l'ensemble des données?

Le choix de la méthode d'extraction des séries de pics sanitaires (ou valeurs extrêmes) a un impact direct sur leur type de distribution. Dans notre projet, nous avons retenu deux distributions théoriques applicables aux valeurs extrêmes : la distribution généralisée de valeurs extrêmes et la distribution généralisée de Pareto (annexe 1).

La distribution généralisée de valeurs extrêmes a été développée à l'aide de la méthode d'extraction par blocs et la distribution généralisée de Pareto, avec la méthode de dépassements par seuil. Nous les présentons ci-après, suivies d'un exemple tiré de nos résultats.

Extraction des séries de pics avec les blocs

Dans notre projet, les pics des issues sanitaires pour cause de maladies cardiovasculaires sont les maxima d'une loi inconnue. Cette loi peut toutefois être approximée par la loi généralisée des valeurs extrêmes (Fisher & Tippett, 1928), lorsque la taille de leur échantillon devient grande (il s'agit ainsi d'une loi limite). Ce faisant, les pics sont extraits par « bloc ».

Concrètement, il s'agit de séparer la série temporelle de données brutes en plusieurs blocs de même taille, desquels seule l'observation maximale (ou pic) est conservée. Il y a donc autant de pics que de blocs. Ces pics constituent la série à analyser; leur distribution limite est une distribution généralisée de valeurs extrêmes.

Les blocs peuvent être constitués selon les saisons ou un nombre fini d'observations, ou selon les besoins des organismes concernés. Il n'existe pas de procédure formelle pour déterminer le nombre de blocs à retenir. Ce choix est toutefois important puisqu'il influe directement sur les caractéristiques de la distribution des pics (Gilli & Këllezi, 2006; McNeil & Frey, 2000). En pratique, on sait toutefois que l'extraction d'un faible nombre entrainera une courte série d'extrêmes, ce qui entraînera une variance élevée. En contrepartie, un nombre trop élevé aboutira à une longue série d'extrêmes, risquant ainsi de sélectionner des évènements non extrêmes, ce qui peut introduire un biais important.

En l'absence de méthode pour déterminer les blocs à extraire et de publications dans le domaine de la santé cardiovasculaire y faisant appel, nous avons exploré les blocs de 7, 14, 21, 30, 60, 90, 120 et 180 jours.

Extraction des séries de pics à l'aide des seuils

La distribution généralisée de Pareto a été introduite par Smith (1987). Elle repose sur l'établissement d'un seuil élevé relativement aux données brutes (quantile à 90 %, etc.). Une fois le seuil fixé, toutes les observations qui se situent au-delà sont considérées comme étant des pics, d'où le nom de méthode des dépassements du seuil (ou *peaks-over-threshold (POT)*). Ces dépassements peuvent alors être modélisés grâce à la distribution de Pareto.

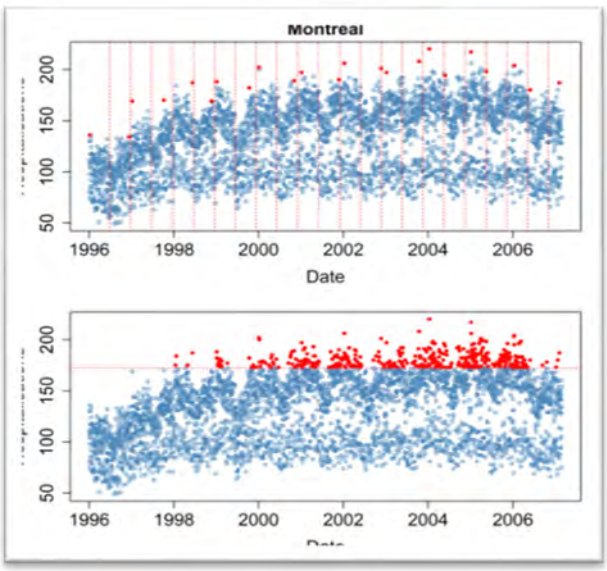
Il n'existe pas de procédure pour déterminer le seuil u le plus approprié, bien que certains outils statistiques puissent soutenir la prise de décision à ce sujet (Lang, Ouarda & Bobée, 1999). Toutefois, comme pour la méthode par bloc, ce choix doit conduire à un équilibre entre la précision et la validité des paramètres estimés (Reiss & Thomas, 2007). Ainsi, un seuil trop élevé donnera un nombre faible d'extrêmes, ce qui engendre une variance élevée, tandis qu'un seuil trop faible risque de sélectionner des évènements non extrêmes, ce qui peut biaiser les estimations.

Des études citées précédemment, De Zea Bermudez et Mendes (2012) ont utilisé un quantile à 90 % comme seuil standard, alors que Watts et collab., (2006) ont considéré les quantiles à 80, 85, 90 et 95 %. Pour notre part, nous avons exploré les quantiles de 75, 80, 85, 90, 92,5, 95, 97,5 et 99 %.

Exemple d'extraction des pics par blocs et par seuils

La figure 1 montre les séries de pics (en rouge) extraits des 4 077 jours d'hospitalisation pour cause principale de MCV (nommées hospitalisations ci-après) de la RMR de Montréal de janvier 1996 à mars 2007. Le graphique du haut présente les résultats d'une extraction par blocs et celui du bas, ceux d'une extraction par seuil (fixé à 90 %). On remarque que davantage de pics sont extraits avec la méthode par seuil.

Figure 1 Illustration des pics d'hospitalisations dans la RMR de Montréal, 1996-2007, méthode par bloc (haut) et seuil (bas).



Comment estimer les paramètres des distributions de pics?

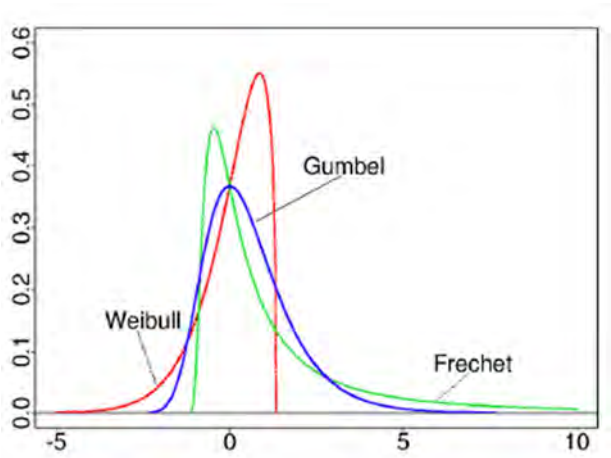
Une fois les séries de pics extraites, nous devons estimer les paramètres de leurs distributions, soit la distribution généralisée des valeurs extrêmes et la distribution généralisée de Pareto dans notre projet (voir annexe 1). Cette étape est importante, car ces paramètres caractérisent des séries qui sont loin de se comporter normalement (courbe gaussienne). En outre, la forme de la distribution change selon le nombre de blocs extraits ou le seuil retenu (figure 2-3, pour un exemple avec la méthode par seuil).

Figure 2 Distribution de pics d'hospitalisations selon des seuils de 75 à 99 %, RMR de Montréal, 1996-2007

Note : *Threshold* signifie Seuil.

Toutefois, quelle que soit leur distribution originale (figure 3, pour des exemples), les pics sont toujours attirés vers une distribution particulière de valeurs extrêmes. Par exemple, les maxima d'une distribution uniforme ou d'une distribution Beta ont tendance à suivre une distribution de Weibull.

Figure 3 Distributions de Weibull, Gumbel et Frechet



Source : Google.

Le champ d'études est aussi important à considérer. Ainsi, les vents extrêmes sont souvent exprimés à l'aide d'une distribution de Weibull, tandis que la distribution de Gumbel est généralement utilisée pour les précipitations extrêmes. Il est toutefois nécessaire que le champ d'études ait exploité la théorie des valeurs extrêmes, ce qui est rarissime dans le cas de la recherche en santé, comme déjà mentionné. Ainsi, davantage d'études sont nécessaires pour obtenir des résultats sur les distributions applicables aux pics dans le domaine de la santé.

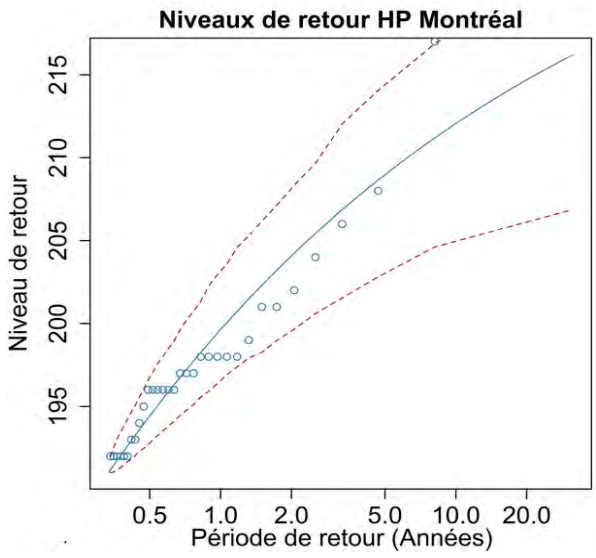
Enfin, deux méthodes ont permis l'estimation des paramètres. Il s'agit de la méthode du maximum de vraisemblance (ou *maximum likelihood*), souvent utilisée dans le domaine de la santé, et la méthode des L-moments, qui démontre de meilleures performances avec de petits échantillons (pour une comparaison des deux méthodes (Hosking, 1990). Nous avons aussi observé la meilleure performance des L-moments dans notre étude.

Comment calculer le niveau de retour?

Une fois la distribution appropriée sélectionnée et les paramètres associés estimés, le niveau de retour est calculé. Rappelons qu'il s'agit du niveau qui sera atteint ou dépassé en moyenne une fois, sur une certaine période, et qu'il peut être vu comme un quantile (nommé aussi « *value-at-risk* » en finance).

La Figure 4 montre le niveau de retour pour le modèle global retenu dans notre étude. Elle indique qu'on s'attend à dépasser 198 hospitalisations par jour 1 fois en 1 année, tandis que sur une période de 10 ans, le pic maximal attendu est de 212. Il s'agit d'une augmentation des pics de 7 % sur 10 ans (comparativement à 1 an). Comme avec toute technique de prévision, l'incertitude traduite par les intervalles de confiance (lignes rouges pointillées) croît rapidement avec le temps.

Figure 4 Courbe du niveau de retour des pics d'hospitalisations, RMR de Montréal, 1996-2007



Dans les graphiques correspondants au niveau de retour, l'abscisse qui représente le temps est couramment tracée selon une échelle logarithmique (Coles, 2001).

Les équations pour calculer le niveau de retour pour les distributions généralisées des valeurs extrêmes et de Pareto sont présentées à l'annexe 1. Leurs intervalles de confiance ont été calculés selon la méthode Delta (Coles, 2001).

Est-il nécessaire de prétraiter la série de pics?

Dans le cas de l'approche classique de la théorie des valeurs extrêmes, les hypothèses d'indépendance et de distribution identique des pics sanitaires doivent être vérifiées avant de procéder à leur modélisation par une distribution extrême (Coles, 2001). De façon équivalente, la stationnarité, l'indépendance et l'homogénéité des pics sanitaires doivent être testées. Dans notre étude, les tests suivants ont été utilisés à cette fin (annexe 2) :

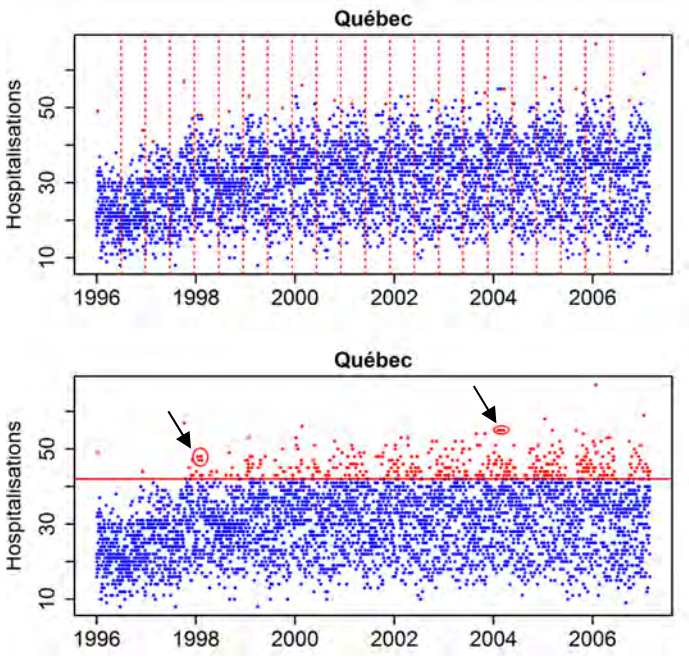
- Mann-Kendall pour l'hypothèse nulle (H0) : les données ne présentent pas de tendance (série stationnaire);
- Wald-Wolfowitz pour l'hypothèse nulle (H0): les données sont indépendantes;
- Wilcoxon pour l'hypothèse nulle (H0): les données sont homogènes[3].

Par ailleurs, les pics ont tendance à survenir en agrégats (ou *clusters*), ce qui peut compromettre l'hypothèse d'indépendance des pics. Cela survient en particulier avec la méthode de dépassements du seuil, mais généralement pas avec la méthode des blocs, sauf lorsqu'ils ont un pas temporel très court (comme des blocs journaliers). Par exemple, à la figure 5, on observe deux agrégats de pics d'hospitalisations (encerclés en rouge, en 1998 et en 2004) dans la RMR de Québec et ceci est présent seulement avec la méthode par seuil (fixé à 90 %).

[3] Si ces conditions ne sont pas respectées, il est possible de considérer une modélisation des paramètres des distributions à l'aide de covariables ou l'utilisation de la régression quantile.

Note. L'extraction de la série de pics a été effectuée avec la méthode de dépassements du seuil, pour un seuil fixé à 95 %. La distribution de la série de pics est donc la distribution généralisée de Pareto.

Figure 5 **Pics d'hospitalisations dans la RMR de Québec, 1996-2007, selon les méthodes par blocs (figure du haut) et de dépassements du seuil fixé à 90 % (figure du bas)**



Afin de s'assurer de l'indépendance des données, il s'avère nécessaire de « déclusteriser » la série de pics (Beirlant et collab., 2004).

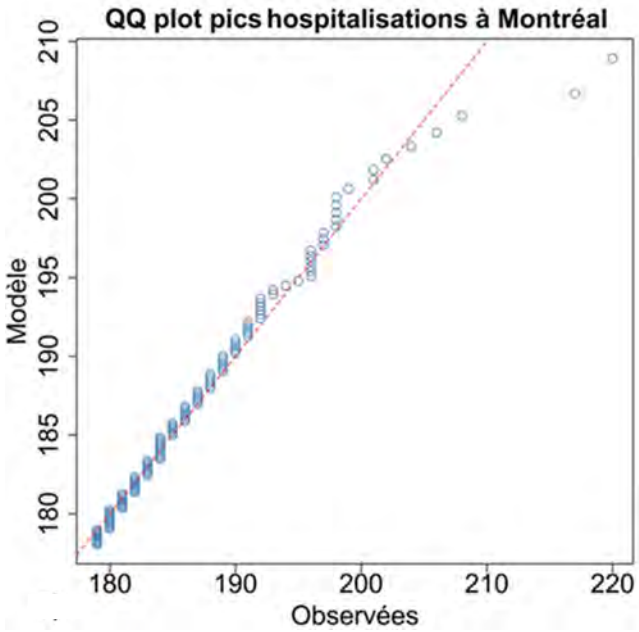
L'ajustement des distributions théoriques est-il adéquat?

Divers tests permettent de vérifier si les distributions théoriques retenues (dans ce cas-ci, les distributions généralisées des valeurs extrêmes et de Pareto) représentent bien la série de données observées. Dans notre étude, nous avons utilisé les tests d'Anderson-Darling et de Kolmogorov-Smirnov. Ceux-ci sont basés sur la distance entre les distributions théoriques et empiriques. Tous deux font partie des tests d'adéquations non paramétriques les plus utilisés pour les extrêmes (Stephens, 1977 et Laio, 2004).

L'hypothèse nulle (H0) est la suivante : la série de pics suit une distribution généralisée des valeurs extrêmes (ou de Pareto), selon les paramètres estimés. Le seuil statistique (α) retenu dépend de l'objet d'étude. Nous avons choisi un α de 5 % puisque nous étions dans un mode exploratoire.

La figure 6 présente le diagramme quantile-quantile (QQ) de l'ajustement du modèle global retenu dans notre étude. Les observations (cercles bleus) sont alignées sur la première bissectrice (ligne rouge pointillée), indiquant une bonne adéquation avec les quantiles théoriques. Les deux plus grandes observations, égales à 217 et à 220, sont tout de même écartées de la droite. Ces valeurs sont vraisemblablement sous-estimées par le modèle, qui suggère plutôt les valeurs 206 et 209.

Figure 6 **Diagramme QQ des pics d'hospitalisations, RMR de Montréal, 1996-2007**



Note. L'extraction de la série de pics a été effectuée avec la méthode de dépassements du seuil, pour un seuil fixé à 95 %. La distribution de la série de pics est donc la distribution généralisée de Pareto.

Le choix des distributions théoriques est-il adéquat?

La distribution généralisée des valeurs extrêmes et la distribution généralisée de Pareto sont des distributions théoriques (asymptotiques). Ainsi, il s'avère fondamental de vérifier l'adéquation des pics sanitaires à d'autres distributions extrêmes mieux connues. Pour ce faire, nous avons retenu la distribution exponentielle, la distribution lognormale et la distribution gamma, utilisées notamment en hydrologie et en finance (Engeland, Hisdal & Frigess, 2004). Ces distributions ont été comparées entre elles à l'aide de la racine carrée de la moyenne des erreurs quadratiques (Gomes & Guillou, 2014). Son équation est présentée à l'annexe 1.

Au final, quel modèle retenir?

Le choix du modèle à retenir dépend de divers critères, qu'ils soient statistiques, cliniques ou organisationnels.

Nous avons opté pour des critères statistiques, car notre étude avait une visée méthodologique. Ainsi, pour qu'un modèle (et, implicitement, une taille de bloc ou un seuil) soit retenu, il devait d'abord être considéré comme étant valide. Cela signifie que la série des pics ne devait être rejetée par aucun des tests d'hypothèses et par aucun des tests d'ajustement. Ensuite, parmi les modèles jugés valides, on retenait celui qui possédait la racine carrée de la moyenne des erreurs quadratiques la plus faible.

Le tableau 1 illustre ces propos pour la série de pics d'hospitalisations dans la RMR de Montréal, selon la méthode de dépassements des seuils. On y note que les valeurs p des tests permettent d'accepter les hypothèses pour les seuils de 95 %, 97,5 % et 99 % (MK, WW et WK, 2e colonne). Peu importe le seuil parmi les trois, les valeurs p des tests (KS et AD, 4e colonne) sur l'ajustement de la distribution généralisée de Pareto (GPD) permettent d'accepter l'hypothèse nulle; les autres distributions (données non présentées) performent moins bien. Par ailleurs, de ces trois seuils, celui de 95 % a les plus faibles valeurs RMSE).

Tableau 1 Résultats de la modélisation des pics d'hospitalisations dans la RMR de Montréal, 1996-2007

MONTRÉAL															
GPD		**Hypothèses (valeur p)**			**Paramètres estimés**		**Adéquation (valeur p)**		**Distributions alternatives (valeur p)**			**Comparaison (RMSE)**			
Seuil	**n**	**MK**	**WW**	**WX**	$\hat{\xi}$	$\hat{\sigma}$	**KS**	**AD**	**EXP**	**LNO**	**GAM**	**GPD**	**EXP**	**LNO**	**GAM**
0,75	269	0	0	0,25	-0,96	36,12	0,02	0,08	0	0,12	0,1	2,95	164,09	1,86	2,01
0,80	280	0	0	0,04	-0,82	29,04	0,05	0,11	0	0,01	0,01	2,56	166,49	2,03	2,16
0,85	257	0	0	0,04	-0,57	20,34	0,33	0,24	0	0,01	0,01	1,9	169,2	2,37	2,48
0,90	218	0,03	0,14	0,01	-0,3	13,55	0,21	0,31	0	0	0	1,14	172,11	2,62	2,73
0,925	187	0	0,08	0,03	-0,31	13,25	0,26	0,31	0	0,01	0	1,16	173,8	2,54	2,64
0,95	**133**	**0,13**	**0,17**	**0,24**	**-0,37**	**13,**14	**0,45**	**0,35**	**0**	**0,03**	**0,02**	**1,**48	**177,04**	**2,49**	**2,57**
0,975	77	0,31	0,57	0,04	-0,32	10,89	0,58	0,36	0	0,04	0,04	1,61	180,58	2,62	2,7
0,99	33	0,29	0,95	0,93	-0,23	8,89	0,45	0,33	0	0,05	0,05	1,78	181,85	2,91	3

Légende. GPD : distribution généralisée de Pareto. RMSE : erreurs quadratiques moyennes. N : taille de l'échantillon. MK : test de Mann-Kendall. WW : test de Wald-Wolfowitz. WX : test de Wilcoxon. $\hat{\xi}$: Paramètre de forme. $\hat{\sigma}$: Paramètre d'échelle. KS : test de Kolmogorov-Smirnov. AD : test d'Anderson-Darling. EXP : distribution exponentielle. LNO : distribcarution lognormale. GAM : distribution Gamma.

Conclusion

Dans le domaine de la santé, les pics sanitaires sont rarement étudiés et ont essentiellement été assimilés à la valeur moyenne des observations, réduisant ainsi leur expression singulière et leur portée. En fait, à ce jour, les pics ont été essentiellement assimilés à la valeur moyenne des observations, réduisant ainsi leur expression singulière et leur portée.

Ce document propose une démarche scientifique et riche d'enseignements incluant non seulement les étapes d'identification, d'extraction et de modélisation des pics, mais aussi diverses techniques statistiques applicables à ces valeurs extrêmes. En outre, il suggère un outil diagnostic intéressant et pouvant être vu comme un quantile : le niveau de retour.

Depuis des décennies, l'ensemble des théories et des techniques décrites dans ce document sont notamment utilisées avec succès dans les domaines de l'hydrologie et de la finance. Le domaine de la santé a donc tout avantage à commencer à se les approprier, car on y trouve là aussi des valeurs extrêmes. L'étude des pics sanitaires avec les outils statistiques adéquats pourrait donc clarifier certains résultats difficiles à interpréter et ainsi soutenir les gestionnaires pour une prise de décision plus éclairée. Les pics sanitaires, aussi nommés « éclosions » en maladies infectieuses, constituent un défi de surveillance majeur et tout spécialement dans un contexte de changements climatiques.

Références

AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., & Sorooshian, S. (Eds.). (2013). Extremes in a changing climate: detection, analysis and uncertainty (Vol. 65). Springer Science & Business Media.

Bayentin, L., El Adlouni, S., Ouarda, T. B., Gosselin, P., Doyon, B., et Chebana, F. (2010). Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. International journal of health geographics, 9(1), 1.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., et Ferro, C. (2004). *Statistics of Extremes : Theory and Applications*: Wiley, Chichester.

Chen, J., Lei, X., Zhang, L., et Peng, B. (2015). Using Extreme Value Theory Approaches to Forecast the Probability of Outbreak of Highly Pathogenic Influenza in Zhejiang, China. *PloS one* no. 10 (2).

Chiu, Y., Chebana, F., Abdous, B., Bélanger, D., et Gosselin, P. (2015). Modélisation des pics de mortalité et de morbidité hospitalière pour cause de maladies cardiovasculaires à Québec et Montréal (Québec) : Une approche par la théorie des valeurs extrêmes. Version finale. Rapport de recherche (R1593). INRS, Centre Eau Terre Environnement, Québec.

Chiu, Y., Chebana, F., Abdous, B., Bélanger, D., et Gosselin, P. (2016). Mortality and morbidity peaks modeling : An extreme value theory approach. Statistical Methods in Medical Research, 0962280216662494.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Vol. 208: Springer.

Daigle, J.-M. (2007). *Les maladies du coeur et les maladies vasculaires cérébrales : Prévalence, morbidité et mortalité au Québec*: Direction planification, recherche et innovation, Unité connaissance-surveillance, Institut national de santé publique Québec.

De Zea Bermudez, P., et Mendes, Z. (2012). Extreme value theory in medical sciences : Modeling total high cholesterol levels. *Journal of Statistical Theory and Practice* no. 6 (3):468-491.

El Adlouni, S., Chebana, F., & Bobée, B. (2009). Generalized extreme value versus Halphen system : Exploratory study. Journal of Hydrologic Engineering, 15(2), 79-89.

Engeland, K., Hisdal, H., et Frigessi, A. (2004). Practical extreme value modelling of hydrological floods and droughts : a case study. *Extremes* no. 7 (1):5-30.

Fisher, R. A., et Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Paper read at Mathematical Proceedings of the Cambridge Philosophical Society.

Gilli, M. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics* no. 27 (2-3):207-228.

Girard, C., Binette Charbonneau A., Payeur F. F., (2016). *Le bilan démographique du Québec*: Institut de la statistique du Québec.

Gomes, M. I., et Guillou, A. (2014). Extreme Value Theory and Statistics of Univariate Extremes: A Review. *International Statistical Review*:n/a-n/a. doi: 10.1111/insr.12058.

Guillou, A., Kratz, M., et Le Strat, Y. (2014). An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella. *Stat Med* no. 33 (28):5015-27. doi: 10.1002/sim.6275.

Hosking, J. R. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*:105-124.

Laio, F. (2004). Cramer–von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research* no. 40 (9).

Lang, M., Ouarda, T. B. M. J., et Bobee, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology* no. 225 (3-4):103-117. doi: Doi 10.1016/S0022-1694(99)00167-5.

Longin, F. M. (2000). From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance* no. 24 (7):1097-1130. doi: Doi 10.1016/S0378-4266(99)00077-1.

McNeil, A. J., et Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance* no. 7 (3):271-300.Reiss, R., et Thomas, M. 2007. Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields.: Birkhauser.

Reiss, R. D., et Thomas, M. (2007). Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields. Birkhäuser Basel. Springer. ISBN: 978-3-7643-5768-9 (Print) 978-3-0348-6336-0.

Smith, R. L. (1987). Estimating tails of probability distributions. The annals of Statistics, 1174-1207.

Salvadori, G., De Michele, C., et Durante, F. (2011). On the return period and design in a multivariate framework. Hydrology and Earth System Sciences no. 15 (11):3293-3305. doi: 10.5194/hess-15-3293-2011.

Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. Biometrika no. 64 (3):583-588.

Watts, K. A., Dupuis, D. J., et Jones, B. L. (2006). An extreme value analysis of advanced age mortality data. North American Actuarial Journal no. 10 (4):162-178.

Annexe 1 Équations

1. Estimation des paramètres

Distribution généralisée des valeurs extrêmes	Distribution généralisée des Pareto
Soit *X* la série de pics extraits, la fonction de la distribution généralisée des valeurs extrêmes se définit par cette équation (1) : $G\left(x;\mu,\sigma,\xi\right)=\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$ pour $\xi\neq 0$ pour $\left\{x:1+\xi\left(\frac{x-\mu}{\sigma}\right)>0\right\}$, où $-\infty<\mu<\infty$ est le paramètre de position, $\sigma>0$ le paramètre d'échelle et $-\infty<\xi<\infty$ le paramètre de forme. Dans le cas spécial $\xi=0$, l'équation 1 est réduite à $G\left(x;\mu,\sigma,\xi\right)=\exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}$. La distribution de Weibull, de Fréchet ou de Gumbel est obtenue si $\xi<0$, $\xi>0$ ou $\xi=0$, respectivement[A].	Soit *X* la variable de la série brute, les dépassements de *X* au-delà d'un seuil *u* sont représentés par $Y=X-u$. La fonction de la distribution généralisée de Pareto (DGP) est définie par cette équation (2) : $H\left(\mathrm{y};\sigma,\xi\right)=1-\left(1+\frac{\xi y}{\sigma}\right)^{-1/\xi}$ pour $\xi\neq 0$ sur l'intervalle $\left\{\mathrm{y}:\mathrm{y}>0 \text{ et } (1+\xi y/\sigma)>0\right\}$, où $\sigma>0$ est le paramètre d'échelle et $-\infty<\xi<\infty$ le paramètre de forme. Dans le cas particulier $\xi=0$, l'équation 2 est réduite à $H\left(\mathrm{y};\sigma,\xi\right)=1-\exp(-\frac{y}{\sigma})$. Selon le signe de ξ, la DGP prend trois formes distinctes. $\xi>0$ implique une distribution sans limites supérieures (distribution Pareto); $\xi<0$, une distribution avec une borne supérieure (distribution de type Beta); et, $\xi=0$, une distribution exponentielle non bornée (AghaKouchak et collab., 2013)[A].

[A] Cette notation du paramètre de forme est conventionnelle dans la littérature statistique, mais d'autres domaines peuvent préférer la forme $\xi^{*}=-\xi$ (par exemple en hydrologie, El Adlouni et collab., 2009).

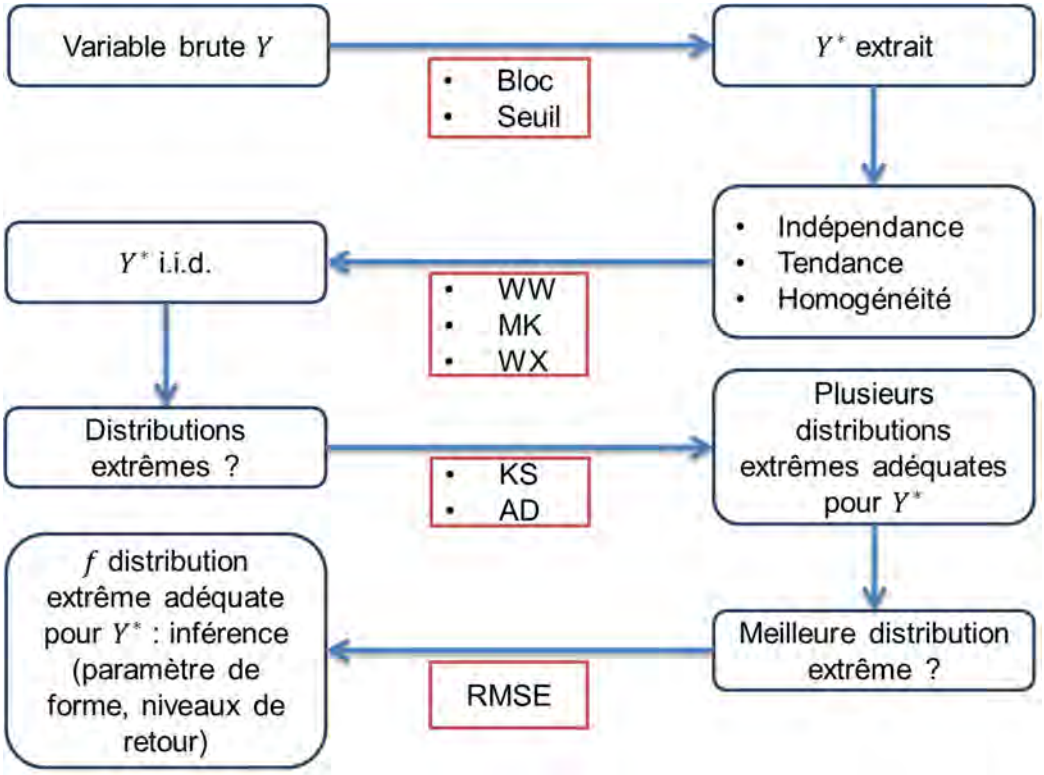
2. Calcul du niveau de retour

Distribution généralisée des valeurs extrêmes	Distribution généralisée des Pareto
Le niveau de retour z_p, lié à une période de retour *T* (équivalent au quantile d'ordre *p* où $T = \frac{1}{p}$, Coles, 2001), se définit par cette équation (3) : $z_p = \mu - \frac{\sigma}{\xi}\left\{1 - \left[-\ln\left(1-p\right)\right]^{-\xi}\right\}$ pour $\xi \neq 0$ $= \mu - \sigma \ln\left[-\ln\left(1-p\right)\right]$ pour $\xi = 0.$	Le niveau de retour z_p de la distribution généralisée de Pareto s'exprime en inversant l'équation (2). Il en résulte cette équation (4) : $z_p = \frac{\sigma}{\xi}\left\{\left(1-p\right)^{-\xi} - 1\right\}$ pour $\xi \neq 0$ $= -\sigma \ln\left(1-p\right)$ pour $\xi = 0.$

3. Comparaison des distributions entre elles

Distribution généralisée des valeurs extrêmes et distribution généralisée des Pareto
Afin de comparer les distributions entre elles, nous avons retenu la racine carrée de la moyenne des erreurs quadratiques (RMSE, Gomes & Guillou, 2014). La RMSE est définie par : $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$

Annexe 2 Diagramme schématisant les étapes de l'étude



Légende. WW : test de Wald-Wolfowitz. MK : test de Mann-Kendall. WX : test de Wilcoxon. KS : test de Kolmogorov-Smirnov. AD : test d'Anderson Darling. RMSE : erreurs quadratiques moyennes.

Modélisation de pics sanitaires à l'aide de la théorie des valeurs extrêmes

AUTEURS

Yohann Chiu[1]
Fateh Chebana[1]
Belkacem Abdous[2]
Diane Bélanger[1,3]
Pierre Gosselin [2,4]

[1] Institut national de la recherche scientifique - Centre Eau Terre Environnement, Québec

[2] Département de médecine sociale et préventive, Faculté de médecine, Université Laval, Québec

[3] Centre de recherche du Centre hospitalier universitaire de Québec, Québec

[4] Institut national de santé publique du Québec, Québec

RELECTURE DU DOCUMENT

Mariève Doucet
Bureau d'information et d'études en santé des populations

MISE EN PAGE

Nabila Haddouche
Bureau d'information et d'études en santé des populations

ORGANISME SUBVENTIONNAIRE

Cette étude a été financée par le Fonds vert dans le cadre de l'Action 21 du Plan d'action sur les changements climatiques 2006-2012 du gouvernement du Québec

Dépôt légal – 2e trimestre 2019
Bibliothèque et Archives nationales du Québec
ISBN : 978-2-550-84178-4 (PDF)

Nº de publication : 2554

Institut national de santé publique Québec