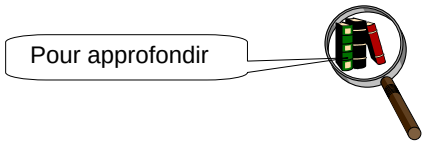

Vigi BN :
La détection des épidémies

Louis AYZAC, C.CLIN Sud Est
Hadrien CHARVAT, Interne Santé Publique

Méthodes de mise en évidence de phénomènes anormaux dans les séries temporelles. Application à la détection des épidémies dans les données de surveillance des bactériémies (Mémoire de Master AMIV Lyon 1 ; 2006-2007 ; Hadrien CHARVAT).

Surveillance épidémiologique : recueil continu et standardisé, analyse, interprétation et diffusion de données relatives à la santé permettant la mise en place de mesures visant à réduire la morbidité et la mortalité et à améliorer l'état de santé de la population source.

Détection de phénomènes anormaux au sein de ces masses de données : **préoccupation émergente** des épidémiologistes et des biostatisticiens, ayant donné lieu à l'élaboration de **multiples techniques**, plus ou moins complexes et efficientes, dont beaucoup sont déjà à l'œuvre au niveau de programmes nationaux de surveillance épidémiologique.

Tempête dans une bouteille de Coca

1970 : une **épidémie nationale** de bactériémies nosocomiales dues à des solutés de perfusion contaminés **échappe à la surveillance du CDC** malgré la présence des données nécessaires à sa découverte(1).

⇒ **Utilité du développement des techniques de détection** des épidémies plus à démontrer.

1. Goldman, D. A., R. E. Dixon, C. C. Fulkerson, D. G. Maki, S. M. Martin, and J. V. Bennett, 1978, The role of nationwide nosocomial infection surveillance in detecting epidemic bacteremia due to contaminated intravenous fluid: Am J Epidem, v. 108, p. 207-213.

Ces méthodes dépendent de la nature et de la qualité des données recueillies

- Elles peuvent concerner la **mise en évidence**
  - d'un phénomène **rare**
  - ou bien encore la **détection d'un nombre élevé d'événements** de même nature
- au sein d'une même unité
  - **spatiale**,
  - **temporelle**
  - ou **spatio-temporelle**.
- Elles trouvent, pour un certain nombre d'entre elles, leur **origine** dans le **milieu industriel**, l'informatique (**data mining** …) ou encore la **finance**.

Limites de l'étude

Méthodes de **détection de phénomènes anormaux au sein d'une ou plusieurs série(s) temporelle(s).**

L'analyse de la **composante spatiale** n'ayant pas de bien-fondé au niveau d'un établissement hospitalier **à moins de pouvoir établir une topologie** rendant compte à la fois des **proximités géographiques** mais aussi des **flux de patients entre les différents services**.

Classement des méthodes de détection selon qu'elles concernent l'analyse <u>rétrospective</u> ou <u>prospective</u> des données.

Dans **l'analyse rétrospective**, la mise en évidence d'une agrégation anormale de cas sera le point de départ d'une **investigation visant à en comprendre** les causes et les moyens d'en éviter la répétition.

Dans **l'analyse prospective**, on cherche à **détecter de façon précoce** la survenue d'un phénomène épidémique afin, si possible, **de l'endiguer ou tout au moins d'en limiter les conséquences** sur la population. Cette approche est actuellement très étudiée, d'une part dans le cadre des **programmes de surveillance des infections** mais surtout devant la menace que représente la possibilité d'un acte de **bioterrorisme**.

Jugement de la qualité de ces méthodes

- en fonction de leurs
  - **Sensibilité**
  - **Spécificité**
- mais aussi de leur
  - capacité à **détecter** le phénomène **le plus rapidement possible**.

- Un ensemble d'outils voit le jour, permettant d'analyser ces paramètres en se fondant sur des **données réelles** (historiques) ou **simulées**.

I - La détection de clusters temporels

- Principales méthodes utilisées pour la détection d'agrégats de cas au sein d'une ou plusieurs série(s) temporelle(s), la plupart ayant été utilisées dans le cadre d'analyses **rétrospectives** mais pouvant être étendues à la détection **prospective** d'anomalies.
- **Augmentation de la fréquence de survenue** de l'événement considéré, la détection de celle-ci se fondant, suivant les techniques, sur l'étude des **intervalles de temps** séparant les cas successifs ou sur le **nombre de cas survenant** au sein d'intervalles de temps déterminés.
- Reposent sur la comparaison de la valeur d'un indicateur (**test**) extraite à partir des **données** à celle obtenue sous **l'hypothèse nulle de répartition aléatoire des événements** au sein des intervalles de temps composant la ou les séries étudiée(s) en parallèle.

Empty Cells Test (test des « cellules vides ») [2]

Ce test peut être utilisé pour la détection de clusters d'événements **rares** au sein d'une ou plusieurs séries temporelles.

La série temporelle est découpée en **intervalles de temps égaux consécutifs** ou « **cellules** ».

La statistique de test **E** représente le **nombre de cellules dans lesquelles aucun cas n'est dénombré**. En cas **d'agrégation temporelle** des cas, cette statistique est **supérieure à son estimation sous H0**. En cas de répartition uniforme des cas, elle est inférieure à son estimation sous H0. Ainsi, ce test est sensible à la présence d'un nombre excessif d'intervalles de temps ne comportant aucun cas.

Cette méthode ne peut s'appliquer **qu'à la recherche de clusters d'événements rares**. En effet, elle nécessite la présence de cellules « vides ». Si cela n'est pas le cas, on peut utiliser des approches similaires telles que les méthodes de Dat ou d'Ederer-Myers-Mantel décrites ci-après.

Cette méthode n'est **pas influencée par la taille** des populations de chaque série temporelle mais **elle l'est par des variations de taille** de ces populations au cours du temps.

2. Résumés des différentes méthodes de détection de clusters temporels sur le site de GeoMed :

http ://zappa.nku.edu/~longa/geomed/require/

Test de Larsen [3]

De même que le test des « cellules vides », cette méthode ne s'applique **qu'à la détection de phénomènes rares** au sein d'une ou plusieurs séries temporelles.

La série temporelle est divisée en **au moins 10 intervalles** de temps égaux consécutifs.

On détermine la position de la « **cellule centrale** » définie comme étant la cellule « occupée » de rang médian parmi toutes les cellules occupées. On calcule ensuite la valeur de la statistique de test **K** correspondant à la **somme (en valeur absolue) des distances de chaque cellule occupée à la cellule centrale**. La valeur de K sera inférieure à son estimation sous H0 en cas de présence d'un cluster et supérieure à son estimation en cas de répartition uniforme des cas.

Cette méthode ne permet **pas de faire la différence entre** la présence de **plusieurs clusters ou la répartition uniforme** des cas au sein de la même série.

Dans les cas où l'on dispose de moins de dix intervalles de temps, il faut se tourner vers la méthode des « cellules vides » (s'il existe des cellules vides) ou le test de Dat si le nombre de cas est plus important.

Cette méthode n'est **pas influencée par la taille** des populations de chaque série temporelle mais **elle l'est par des variations de taille** de ces populations au cours du temps.

*3. Larsen, R. J., C. L. Holmes and C. W. Heath (1973). "A statistical test for measuring unimodal clustering : a description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia" Biometrics **29**: 301-9.*

Méthode de Dat (0-1 Matrix Method) [4]

Ce test permet la détection **d'agrégations temporelles** de cas au sein d'une ou de plusieurs séries temporelles simultanées.

Les séries temporelles sont découpées **en 5 à 10 intervalles de temps** égaux. La statistique de test **A** représente le **nombre d'intervalles de temps** comportant un nombre de cas **au moins aussi important que celui estimé sous H0**. En cas d'agrégation temporelle, A sera inférieure à son estimation sous H0.

Cette méthode ne peut être utilisée **que si chaque intervalle de temps comprend au moins deux cas**. Si ce n'est pas le cas, il faut envisager l'utilisation du test des cellules vides.

Ce test est **plus sensible** que la méthode d'Ederer-Myers-Mantel pour la détection de **clusters multiples**.

Cette méthode n'est **pas influencée par la taille** des populations de chaque série temporelle mais elle **l'est par des variations de taille** de ces populations au cours du temps.

4. Dat, M. V. (1982). "Tests for time-space clustering" Ph. D.dissertation, Dept. Of Biostatistics, SPH, University of North Carolina, Chapel Hill, NC.

Méthode d'Ederer-Myers-Mantel [5,6]

Ce test permet la détection **d'agrégations temporelles** de cas au sein de **plusieurs séries temporelles simultanées**.

Les séries temporelles sont découpées en **2 à 5 intervalles de temps** égaux.

La statistique de test **m1** représente le **nombre maximum de cas survenant dans un intervalle de temps au sein d'une série**. En cas d'agrégation temporelle, m1 est supérieure à son estimation sous H0 et inférieure en cas de répartition uniforme des cas.

Cette méthode n'est **pas influencée par la taille** des populations de chaque série temporelle mais elle **l'est par des variations de taille** de ces populations au cours du temps.

*5. Ederer, F., M. H. Myers, et al. (1964). "A statistical problem in space and time : do leukemia cases come in clusters?" Biometrics **20**: 626-638.*

*6. Stark, C. R. And N. Mantel (1967). "Lack of seasonal or temporal spatial clustering of Down's Syndrome births in Michigan" Am J Epidem **86**: 199-213.*

Test de Grimson [7]

Cette méthode peut être utilisée dans la **détection d'agrégat spatiaux, temporels ou spatio-temporels**.

La série temporelle étant divisée en **intervalles consécutifs**, l'expérimentateur étiquette ceux qu'il considère comme « **à risque** » (nombre de cas ou taux supérieur à un seuil fixé). La statistique de test **A** correspond au **décompte du nombre de cellules « à risque »** adjacentes. Elle est comparée à son estimation sous l'hypothèse nulle **d'étiquetage aléatoire** des cellules.

Cette méthode **dépend** entièrement des choix du **nombre de cellules**, du nombre de **celles que l'on étiquette** (donc du seuil de non-normalité choisi) et de la valeur de **A déterminée par l'expérimentateur**.

7. Grimson, R. C. (1989). "Assessing patterns of epidemiologic events in space-time" Proceeding of the 1989 Public Health Conference on Records and Statistics, National Center for Health Statistics.

Scan Test de Naus [8-10]

Dans cette approche, on fait défiler une **fenêtre temporelle de longueur fixe** (déterminée en fonction de paramètres tels que le temps d'incubation de la maladie) le long de la série temporelle et la statistique de test **Sw** représente le **nombre maximum de cas survenant dans un des intervalles** de temps ainsi délimités. Sw est supérieure à son estimation sous H0 en cas de présence d'un cluster et inférieure en cas de répartition uniforme des cas.

Cette méthode n'est **pas influencée par la taille** des populations de chaque série temporelle mais elle **l'est par des variations de taille** de ces populations au cours du temps.

Le principal problème posé par cette méthode réside dans la **détermination du degré de signification p** de la valeur obtenue. En effet, celui-ci dépend du nombre total d'événements N durant la période considérée et de la taille de la fenêtre temporelle. Il existe des **tables pour de faibles valeurs de N** mais l'extension de la méthode à des données plus importantes nécessite le recours à des **formules approchées**, objets de nombreuses publications [11-15].

*8. Naus, J. (1965). "The distribution of the size of the maximum cluster of points on a line." J Am Stat Ass **60**: 532-538.*

*9. Naus, J. (1966). "A power comparison of two tests for non-random clustering." Technometrics **60**: 532-538.*

*10. Wallenstein, S. (1980). "A test for detection of clustering over time." Am J Epidemio **11**(3): 367-372.*

*11. Wallenstein, S. and J. Naus (1974). "Probabilities for the size of largest clusters and smallest intervalls." J Am Stat Ass **69**: 690-697.*

*12. Wallenstein, S. and J. Naus (1973). "Probabilities for the kth nearest neighbor problem on the line." Ann Probab **1**: 188-190.*

*13. Wallenstein, S. (1987). "An approximation for the distribution of the scan statistic." Stat Med **6**: 197-207.*

*14. Wallenstein, S. and J. Naus (2004). "Scan statistics for temporal surveillance for biologic terrorism." Inconnu **53**((suppl)): 74-78.*

15. Loader, C. (Non daté). "Large deviation approximations to the distribution of scan statistics." Inconnu: 1-30.

II – La détection précoce de phénomènes anormaux

On s'intéressera ici à tout un ensemble de méthodes fondées sur la **modélisation des séries temporelles** de cas permettant de fournir un **signal d'alerte**, prélude à des investigations complémentaires, lorsque la valeur enregistrée par le système de recueil des données **dépasse** de façon « anormale » **celle prévue par le modèle**.

La difficulté de ces approches réside dans le choix du **modèle**, la **transformation** adéquate de la série avant modélisation, l'estimation des **paramètres** du modèle et enfin la **mise à jour** du modèle en fonction des données enregistrées. En particulier, la détermination des paramètres du modèle s'appuyant sur des données **historiques** plus ou moins anciennes, se pose le problème de **différences** dans la **définition** des cas ou dans les **méthodes de diagnostic** utilisées au cours du temps. Par ailleurs, il faut pouvoir **minimiser** le poids relatif des **phénomènes épidémiques passés** sous peine d'augmenter artificiellement le seuil de détection d'une anomalie future.

Méthodes issues du contrôle de la qualité dans le milieu industriel [16,17]

Ces techniques sont issues du **monde industriel**, utilisées initialement dans les processus d'amélioration de la qualité (cartes de contrôle,…). Les plus utilisées dans le domaine de la santé sont celles faisant intervenir les **sommes cumulatives des déviations entre valeurs attendues et observées**. On insistera en particulier sur les techniques de **CuSUM**. Dans sa formulation traditionnelle, cette méthode **compare** la **valeur observée** à une **moyenne théorique constante au cours du temps**. Ce modèle a été raffiné pour tenir compte de la nature périodique des phénomènes de santé (corrections tenant compte des **variations saisonnières et journalières**). Il est par exemple mis en application dans le système **EARS** (Early Aberration Reporting System) du **CDC**.

*16. Hutwagner, L. C., T. Browne, et al. (2005). "Comparing aberration detection methods with simulated data." Emerging Infect Dis **11**(2): 314-31*

*17. Hutwagner, L. C., E. K. Maloney, et al. (1997). "Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks." Emerging Infect Dis **3**(3): 395-400*

Méthode de Serfling [18,19]

Elle est fondée sur un **modèle de régression cyclique**. Celui-ci comporte une composante **linéaire** décrivant la **tendance séculaire** et une composante **sinusoïdale** représentant les **variations périodiques** du phénomène étudié. Les paramètres sont estimés par la méthode des moindres carrés. Cette méthode est utilisée dans la détection d'épidémies de maladies à **évolution cyclique** telles que la **grippe**.

*18. Serfling, R.E. (1963) "Methods for current statistical analysis of excess pneumonia-influenza deaths" Public Health Reports **78**: 494-506*

19. Tsui, F. C. R., M. Wagner, V. Dato, H. C. Chang (2001). "Value of ICD-9-coded chief complaints for detection of epidemics" Symposium of Journal of American Medical Informatics Association

Modèles de type ARIMA [20,21]

L'idée est ici de modéliser la série en exprimant chacun de ses termes en fonction des valeurs qui le précèdent (**auto-corrélation**). On s'affranchit de la **tendance** (pour obtenir la **stationnarité**) en remplaçant la série initiale par la **série des différences adjacentes**. Une série ainsi différenciée est considérée comme la version « intégrée » (**I** pour **Integrated**) d'une série stationnaire. Les termes de la série sont par ailleurs représentés comme une combinaison linéaire des valeurs les plus récentes (**AR** pour **Auto-Regressive**) plus une composante aléatoire. Cette dernière est elle-même représentée par la combinaison linéaire d'un certain nombre d'erreurs aléatoires antérieures (**MA** pour **Moving Average** : Moyenne Mobile). La prise en compte de la composante saisonnière aboutit à un modèle plus complexe appelé SARIMA.

20. Box, G. E. P., and G. M. Jenkins (1976). "Time series analysis : forecasting and control" San Francisco, CA: Holden-Day, 2nd edition.

*21. Allard, R. (1998). "Use of time-series analysis in infectious disease surveillance" Bulletin of the World Health Organization **76**(4): 327-33*

Algorithme de Farrington [22-24]

Cette méthode est utilisée pour la détection d'événements inhabituels au sein des données de **surveillance hebdomadaires** de certaines maladies infectieuses par le **CDSC** (Communicable Disease Surveillance Centre) en Grande-Bretagne. Sa conception est sous-tendue par la volonté de trouver un algorithme **robuste** applicable à **tous les micro-organismes** enregistrés à l'inverse des techniques précédentes nécessitant une modélisation différente pour chaque série temporelle.

En pratique, la méthode fait appel à un modèle de régression **log-linéaire** prenant en compte la tendance, le cycle saisonnier et les épidémies antérieures, les données étant supposées être distribuées suivant une loi de Poisson. La valeur calculée pour la semaine d'intérêt est analysée en fonction des données obtenues durant la même semaine et celles qui l'entourent sur les cinq années précédentes.

*22. Farrington, C. P., N. J. Andrews, et al. (1996). "A statistical algorithm for the early detection of outbreaks of infectious disease." J R Stat Soc **159**(3): 547-563.*

*23. Farrington, C. P. and A. D. Beale (1993). "Computer-aided detection of temporal cluster of organisms reported to the communicable disease surveillance center." Communicable Disease Report **3**(6): R78-R82.*

*24. "Veille sanitaire : nouveau système, nouveaux enjeux" BEH 2005 **27-28**: 133-140.*

De conception plus délicate, on note aussi l'existence des méthodes suivantes :

- Algorithme de Stroup.
- Transformation de la série par la méthode des ondelettes,
- Utilisation d'un modèle univarié à chaînes de Markov

III – Analyse des différentes méthodes
Place de la sensibilité, de la spécificité et de la précocité

L'importance respective de ces paramètres dépend du but de la méthode [23].

En effet, un système utilisé

–dans la détection d'épidémies de **toxi-infections alimentaires** se devra d'être assez **sensible** pour détecter de petits agrégats

–tandis qu'une méthode ayant pour but la détection d'une attaque **bioterroriste** devra faire preuve d'une grande **rapidité** de mise en évidence de l'anomalie mais aussi d'une **spécificité** suffisante pour éviter de fausses alarmes, sources de panique, de coûts d'investigation et de prévention lourds et de perte de confiance dans le système.

*23. Farrington, C. P. and A. D. Beale (1993). "Computer-aided detection of temporal cluster of organisms reported to the communicable disease surveillance center." Communicable Disease Report **3**(6): R78-R82.*

Données nécessaires

Par ailleurs, l'analyse nécessite de disposer de données réelles portant sur des **populations assez importantes** pour avoir déjà fait l'objet de phénomènes épidémiques ou bien de données obtenues par **simulation**, ce qui pose le problème de la **modélisation** des épidémies.

Les stratégies à l'étude pour améliorer la qualité des algorithmes de détection incluent une meilleure modélisation des séries temporelles fondée sur une **utilisation optimale** de l'information disponible, la détermination de **longueurs des fenêtres** temporelles ou le **choix de seuils** adaptés à l'histoire naturelle de la pathologie considérée (durée d'incubation, forme du signal épidémique…)

Amélioration de la qualité des données

D'autre part, l'effort porte également sur l'amélioration de la qualité des données enregistrées (meilleure **exhaustivité**, **minimisation des délais de signalement et d'analyse des données**) et l'enregistrement de **paramètres supplémentaires** pertinents au regard de la situation à analyser (absentéisme, augmentation des achats en pharmacie, requêtes Internet concernant les questions de santé…).

IV - Données disponibles par le réseau BN Sud Est

Historique

La surveillance des bactériémies nosocomiales fait partie du programme minimum de surveillance des infections nosocomiales élaboré par le **Comité Technique National des Infections Nosocomiales** en **1992**.

Ainsi, en **1993**, le réseau de surveillance des bactériémies nosocomiales **BN Sud-Est** a été créé par Marie-Reine Mallaret, MCU-PH, responsable du service d'hygiène hospitalière du CHU de **Grenoble**, également responsable de l'antenne Grenoble du CCLIN. Elle en assura la coordination jusqu'en **2003**, date à laquelle elle se retira du réseau. L'antenne Grenoble du CCLIN fut fermée et le réseau rapatrié à Lyon, ce qui ne se déroula pas sans heurts, les données recueillies de **1993 à 2003 ne pouvant être récupérées**.

Par ailleurs, depuis **1994**, les données nationales de surveillance ont été progressivement mises en commun dans le cadre de la création d'un rapport **national** par le Réseau d'Alerte, d'Investigation et de Surveillance des Infections Nosocomiales (RAISIN).

En **2005**, le **RAISIN** décida de **ne plus reconnaître la surveillance continue des bactériémies nosocomiales comme priorité nationale**.

Récupération des données

Le **travail de création d'un fichier de données** portant sur la période 1994 à 2004 a été effectué sur un établissement (CHLS).

La **structure erratique** des données recueillies sur les années 1992 à 2004 **réduit considérablement le nombre de variables réellement exploitables**. Ainsi, par exemple, sur les 1.181.405 caractères recueillis et saisis par l'unité d'Hygiène et Epidémiologie du Centre Hospitalier Lyon Sud, elle conduit à l'impossibilité d'en utiliser 517.873 soit **43.8 %.**

La non récupération des données du réseau de 1993 à 2003 conduira à demander une **communication rétrospective** de leurs données aux différents établissements ayant participé au réseau pendant une période de temps significative. Quelques établissements peuvent être concernés par cette récupération de leurs données historiques (Annonay, Aubagne, Avignon, Bourg en Bresse, Fréjus, …).

V - Question statistique posée
Objectifs

Le but de la présente étude est de **mettre en œuvre** quelques-unes des **techniques** énoncées précédemment sur les **données existantes** du réseau et d'en **comparer les caractéristiques en terme de sensibilité, spécificité et précocité de détection**, dans l'objectif de mettre en place au niveau du réseau BN Sud-Est un **système d'alerte capable de détecter une augmentation d'incidence des bactériémies au niveau d'un établissement**.

Cela implique que l'on ne s'intéressera pas à la détection d'une éventuelle **composante spatiale** des phénomènes anormaux, fondée sur une **vision globale** (et non par établissement) du réseau et pour laquelle une **nouvelle étude** devrait être entreprise.

V - Question statistique posée
Données utilisées

L'utilisation des données historiques du réseau (tout au moins celles qui ont pu être récupérées) permettra de tester à la fois les méthodes de détection d'agrégats temporels (analyse rétrospective) et les méthodes de détection précoce d'anomalies (analyse prospective). Il est par ailleurs envisagé de recourir à des données simulées pour répondre au même objectif.

V - Question statistique posée
Confirmation

La comparaison des spécificités et sensibilités nécessite de disposer de **moyens permettant de confirmer**, une fois le signal donné, la présence d'un véritable phénomène épidémique.

On discutera ici l'éventuelle place du **typage des micro-organismes** par le laboratoire de microbiologie. Cependant, il n'est **pas envisageable** de disposer de telles informations sur les données **rétrospectives**.

V - Question statistique posée
Rapidité

Enfin, la précocité de détection de phénomènes anormaux soulève le problème de la **rapidité** du recueil des informations.

Ceci pourra constituer un début de réflexion sur les **conditions de recueil, de saisie et d'analyse** des données nécessaires à la surveillance.