Réseau d'écotoxicologie terrestre et aquatique

Fiche thématique N°14 - Avril 2018

Comment obtenir des résultats significatifs dans vos études écotoxicologiques : un guide pratique

Résumé

La variabilité intrinsèque liée à pratiquement toutes les mesures effectuées en biologie ou en écotoxicologie conduit les chercheurs à effectuer des tests statistiques afin d'évaluer la robustesse de leurs résultats. On publiera d'autant plus facilement et dans un meilleur journal que les tests statistiques donneront un résultat significatif. Dans le cas contraire, il reste bien sûr la possibilité de publier dans Journal of Articles in Support of the Null Hypothesis, Journal of Negative Results in Biomedicine ou Journal of Negative Results, mais la publication de résultats significatifs dans Environmental Health Perspectives est tout de même plus gratifiante. Dans toute recherche, il existe donc la tentation d'obtenir coûte que coûte des effets significatifs. Pour obtenir le niveau souhaité de significativité, différentes méthodes peuvent être utilisées plus ou moins consciemment. Elles sont ici décrites et illustrées par des exemples pour mieux apprendre à s'en méfier.

Introduction

L'écotoxicologie est un domaine des sciences qui cherche à établir les relations, au sens large, entre des contaminants physiques, chimiques ou biologiques et des organismes vivants ou des fonctions de ces organismes vivant, fonctions pouvant être analysées à toutes les échelles spatiales, temporelles ou d'organisation.

L'analyse de ces relations utilise l'outil statistique pour rechercher des effets significatifs. Cependant, des difficultés apparaissent rapidement quand on veut utiliser les outils développés depuis une centaine d'années par les statisticiens en raison des particularités des données issues de l'écotoxicologie.

D'abord, la notion de contaminant n'est pas simple à appréhender. Les cas simples correspondent à des molécules fabriquées par l'Homme et qui n'existent pas à l'état naturel. Des cas plus complexes apparaissent avec les molécules dérivées des précédentes qui peuvent apparaître au cours d'un procédé de fabrication. Mais ces molécules dérivées peuvent aussi apparaître comme des produits de dégradation. Tout ceci pour dire que le nombre de contaminants potentiels est infini et certains, qui existent déjà, ne sont pas encore caractérisés. L'Inventaire Européen des Substances Chimiques Commerciales Existantes (EINECS) contient 100 102 substances chimiques (Geiss et al., 1992) mais ne recense bien entendu qu'une fraction des molécules réellement produites, volontairement ou non, par l'Homme. On est donc dans une situation inédite où le nombre de facteurs explicatifs d'un effet peut être quasiment infini. Bien sûr, en pratique, on n'analyse pas l'ensemble des contaminants possibles mais il est relativement facile d'obtenir le dosage de 200 molécules dans un même échantillon.

Une autre particularité de l'écotoxicologie est que les effets des contaminants peuvent toucher de très nombreuses fonctions. On disposera donc de très nombreuses mesures sur les organismes que l'on cherchera à mettre en relation avec des contaminants potentiels.

Une troisième particularité des études actuelles d'écotoxicologie est d'avoir accès à des cohortes de très grandes tailles, notamment en écotoxicologie humaine.

Dans un premier temps, sur la base d'un survol de la littérature écotoxicologique de suivi de cohorte, nous décrirons les recettes applicables pour augmenter les chances de détecter des effets significatifs, au risque bien entendu de multiplier les faux-positifs. Nous détaillerons les problèmes engendrés par

ces méthodes sur la base du changement de paradigme imposé par le *big data* (Fan et al., 2014), puis quelques solutions seront proposées pour limiter la publication de conclusions erronées.

Recettes pour obtenir des effets significatifs

Prenez des individus en nombre suffisamment important, de l'ordre de la centaine (la centaine de milliers si vous avez la chance de travailler sur des cohortes humaines). N'oubliez pas de prendre des informations sur leur milieu de vie, leur histoire et toute information qui vous semble pertinente ou non, mais qui pourra être utile (voir point 3 ci-dessous). Sur chacun de ces individus, dosez des contaminants divers, organochlorés, métaux lourds, HAP… vous avez le choix. Arrangez-vous pour en avoir au moins une vingtaine. Maintenant étudiez des caractéristiques biométriques chez les individus. Cela peut-être des mesures de taille, de masse, de caractéristiques physiologiques ou psychologiques… tout ce qui vous paraît faisable et surtout si c'est très divers.

Avec ces données en main, voyons comment être certain d'obtenir des effets significatifs.

Le principe général est de pratiquer un nombre élevé de tests statistiques pour être sûr que certains parmi eux s'avèreront positifs. Comment faire ? Trois grandes stratégies existent, qui peuvent être couplées bien sûr, et sont décrites ci-après :

1) Combiner les contaminants dans l'analyse de façon à pouvoir augmenter le nombre de tests. Par exemple, vous pourrez inclure les interactions entre contaminants dans l'analyse. Les interactions dans une analyse linéaire se présentent comme un produit entre les concentrations de différents produits. Limitez-vous aux interactions de premier ordre (entre deux produits) sauf si vraiment vous ne trouvez pas d'effets significatifs, ce qui a peu de chance de se produire. Vous pourrez aussi sommer les contaminants pour analyser des classes de produits, soit des classes chimiques soit des regroupements de produits selon la nature de leurs effets suspectés.
2) Combiner les caractéristiques biométriques pour là encore multiplier les tests. Par exemple, si vous avez une longueur et une largeur, multipliez-les pour obtenir une surface et si vous multipliez avec une hauteur, cela vous fait un volume. Donc à partir de 3 mesures linéaires, vous obtiendrez 3 surfaces et un volume, donc 7 mesures. Vous pouvez faire ce genre de traitement pour un peu tout. Par exemple, à partir de dénombrement de tumeurs en 4 catégories de tailles, vous obtiendrez facilement un nombre total de tumeur, une surface et un volume. A partir de concentration de solutés dans le milieu intérieur, vous pouvez combiner les éléments chimiques, par exemple les anions et les cations.
3) Si ces stratégies ne permettent pas d'obtenir au moins un effet significatif (ce qui est peu probable), vous avez aussi la possibilité de retirer des individus du jeu de données. C'est ici qu'il s'avère utile de connaître les caractéristiques des individus : sur cette base, vous pourrez décider de retirer certains d'entre eux de l'analyse, par exemple, ceux qui sont en surpoids, le genre ou ceux qui proviennent d'une localité X. Vous avez aussi la possibilité de créer des sous-ensembles du jeu de données à partir de ces caractéristiques et d'analyser ces sous-ensembles séparément.

Sur la base de ces très nombreux jeux de données, vous allez donc effectuer de très nombreuses analyses cherchant des relations entre vos combinaisons de variables à expliquer et vos combinaisons de variables explicatives, les contaminants. Soyez en certains : vous trouverez des effets significatifs et n'aurez que l'embarras du choix pour trouver de jolies histoires à raconter.

Vous trouvez que la situation décrite est caricaturale ? Et pourtant les procédures décrites ci-dessus sont celles-là même rencontrées dans nombre de publications très sérieuses et sur les résultats desquels des décisions de politique publique pourront s'appuyer. Pour vous en convaincre, analysez cette étude récente qui a eu la faveur d'avoir un compte rendu dans le journal *Le Monde* : Effet de la pollution de l'air pendant la vie embryonnaire sur le développement du cerveau et les troubles

cognitifs (Guxens et al., 2018). Il est difficile de dénombrer les *p-values* de ce papier car toutes ne sont pas montrées, ce qui interdit ici d'utiliser les procédures de correction pour la multiplication de tests statistiques (cf plus bas), mais elles sont très nombreuses. Cette étude est à mettre en relation avec d'autres sur le même sujet. Par exemple, l'effet potentiel de 52 polluants sur 1 variable à expliquer avec la cohorte entière ou séparée selon le genre a été analysé (Braun et al., 2014). Le nombre de tests total est donc de 156 (52x3) et le nombre de polluants pour lesquels un effet significatif au seuil de 0,05 est détecté est de 6. Il est par ailleurs intéressant de noter que le bisphénol A, bien que testé, n'en fait pas partie alors qu'un effet du bisphénol A avait été observé par une procédure similaire dans une étude antérieure sur un jeu de données qui avait été séparé en plusieurs sous-entités pour l'analyse ce qui conduisait là encore à de nombreux tests (Braun et al., 2009).

Ces différents résultats sont exactement ceux attendus s'ils correspondent à des faux-positifs (effets significatifs variables selon les études, en proportion à peu près égales à 5% soit environ 9 quand 156 tests sont effectués, avec des relations parfois positives et parfois négatives). Tout est fait ici pour trouver des résultats positifs et de telles situations sont extrêmement fréquentes dans la littérature.

Qu'est et que n'est pas la *p-value* ?

Dans un test d'hypothèse nulle, l'hypothèse nulle H_0 est clairement explicitée. L'hypothèse H_1 est l'hypothèse alternative et ce sont toutes les hypothèses qui ne sont pas H_0. Quand on fait un test, on cherche si le jeu de données a pu être obtenu sous l'hypothèse H_0. Il faut noter que la réponse est tout le temps "oui" mais avec une probabilité plus ou moins élevée et la "*p-value*" désigne cette probabilité (Fisher, 1934). Par exemple, si p=0,2, cela signifie qu'un jeu de données au moins aussi extrême avait 20% de chance d'être obtenu si H_0 était vraie. Par convention, le seuil de significativité de 5% est souvent adopté. Dans le cas précédent, le résultat est donc qualifié de "non-significatif" au seuil de 5% et on convient de ne pas rejeter H_0. La *p-value* est donc, d'un point de vue formel : prob($x|H_0$), le signe | signifiant "sachant que" ; C'est donc la probabilité d'observer les données x, sachant que H_0 est vrai.

Mais est-ce vraiment ce que l'on veut savoir ? Le problème est que la *p-value* ne nous donne pas la probabilité qui nous intéresse, celle que H_0 soit vraie sachant qu'on a observé les données x : prob($H_0|x$). La *p-value* nous donne la probabilité d'observer les données x sous l'hypothèse H_0 : prob($x|H_0$). Or, en appliquant le théorème de Bayes, on voit bien que ces deux probabilités ne sont pas égales :

$$prob(H_0|x) = prob(H_0)\ prob(x|H_0) / prob(x)$$

En conclusion:

- La *p-value* n'est pas la probabilité que l'hypothèse nulle soit vraie, ni la probabilité que l'hypothèse alternative soit fausse
- La *p-value* n'est pas la probabilité que les données soient dues simplement au hasard.
- La *p-value* n'est pas la probabilité de rejeter à tort l'hypothèse nulle.
- La *p-value* n'est pas la probabilité que l'expérience donne une conclusion différente si elle est reproduite.

La *p-value* peut être utilisée pour donner une indication sur un effet, sans plus (Burnham and Anderson, 2014).

Des effets significatifs en surnombre ou insignifiants

Testons d'abord les résultats obtenus avec la première stratégie (celle qui consiste à multiplier le nombre de facteurs explicatifs testés). Pour cela, générons un jeu de données au hasard tiré d'une distribution normale avec 250 individus, 10 mesures et 20 contaminants. On utilisera des modèles linéaires en testant chacune des 10 mesures avec les 20 contaminants et toutes les interactions de premier ordre. Chacun des effets sera testé et une *p-value* (Voir encadré) sera classiquement calculée pour chaque facteur dans l'analyse.

Le résultat montré sur la figure 1A est satisfaisant puisqu'on détecte de très nombreux effets significatifs : vous avez démontré l'impact de contaminants sur des fonctions biologiques et, en prime, vous mettez en évidence un effet cocktail (notion très à la mode) puisque jamais moins de 50 effets sont observés au sein d'une analyse ! Vous pouvez aussi ne pas analyser les interactions (Figure 1B) ou bien analyser les polluants un par un (Figure 1C) et vous observerez encore suffisamment d'effets significatifs pour raconter quand même une histoire mettant en relation une combinaison de polluants avec des caractéristiques biologiques.

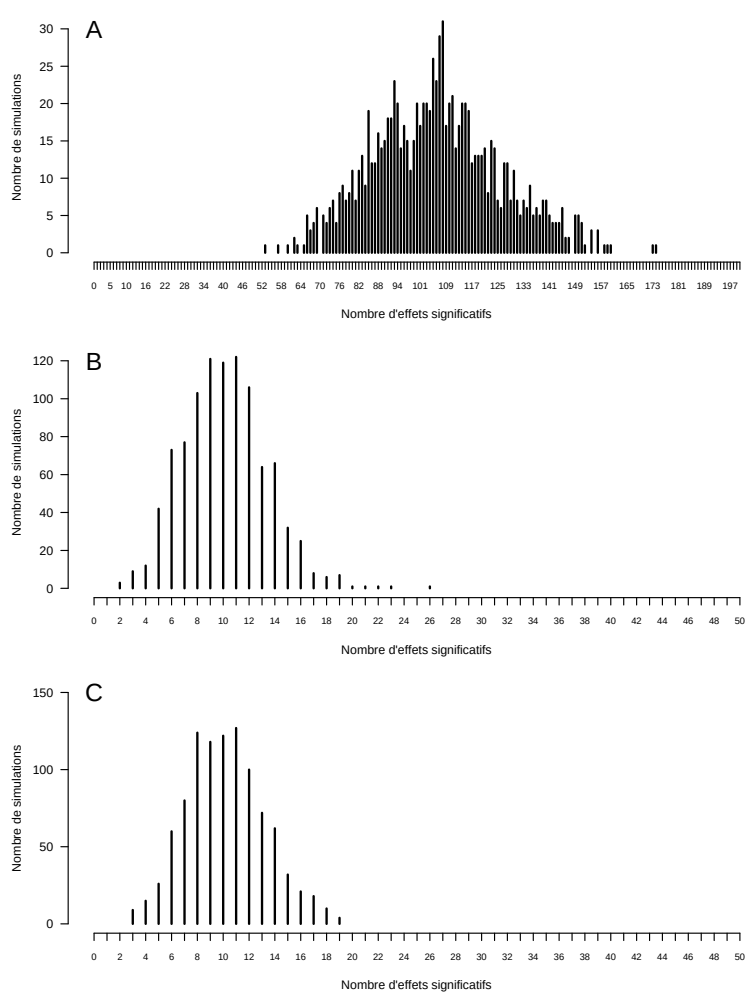

Figure 1: Distribution du nombre d'effets significatifs pour un seuil de 0,05 détectés sur 1000 réplicats avec 10 mesures, 20 contaminants et 250 individus, tous générés au hasard. (A) analyse linéaire avec les 20 contaminants et les 190 interactions de premier ordre testés ensemble, (B) sans les interactions et (C) avec chacun des contaminants testés séparément.

De plus, l'augmentation de la puissance d'un test[1] avec la taille de l'effectif est un phénomène bien connu. Si vous avez la chance de pouvoir utiliser une cohorte de taille importante, vous serez capable de détecter des effets extrêmement faibles, si faibles même que leur signification biologique deviendra hautement hypothétique (Yoccoz, 1991).

Enfin l'utilisation de la *p-value* elle-même peut prêter à confusion puisqu'elle est souvent confondue avec la probabilité de se tromper en concluant à un effet (Voir encadré).

Des éléments de solution

Le calcul de nombreuses *p-values* est le premier problème pouvant biaiser[2] les résultats d'un test statistique. Malheureusement, on voit couramment des papiers d'écotoxicologie avec plus de 100 *p-values* dans le même article sans aucune correction. Les procédures de correction pour les tests multiples sont les corrections de Bonferroni (Dunn, 1959) et la procédure *False Discovery Rate*, moins stringente, (Benjamini and Hochberg, 1995).

Une alternative, illustrée plus haut, consiste à générer des jeux de données purement aléatoires mais avec les mêmes distributions que les observations et à leur appliquer les mêmes méthodes que celles utilisées dans l'étude pour évaluer le nombre de potentiels faux-positifs. Si l'on observe à partir du jeu de données aléatoires un nombre voisin de faux positifs avec des valeurs de *p-value* similaires aux effets significatifs obtenus dans l'article, on peut légitimement se poser la question de la validité de ces derniers. A l'inverse, si les faux positifs sont bien moins nombreux que les effets significatifs de l'article ou avec des *p-values* beaucoup plus faibles, on pourra accorder une certaine crédibilité à certains effets.

Les deux solutions qui viennent d'être proposées ne sont praticables qu'à une condition extrêmement importante : les auteurs doivent préciser de façon exacte le nombre total de tests effectués pour leur étude. Par exemple, supposons qu'une étude incluant 50 tests arrive au résultat d'une unique *p-value*<0,05 mais que l'article ne montre que 5 tests au lieu des 50 effectués, toutes les méthodes de correction et de vérification seront inefficaces. Rapporter tous les tests effectués est donc une absolue nécessité pour que les résultats soient interprétables. Transgresser cette règle de transparence est considéré comme une mauvaise pratique (John et al., 2012).

Il est à noter qu'aucune des deux approches citées précédemment ne s'affranchit des problèmes posés par l'utilisation des *p-values*. Éviter les interprétations erronées de la *p-value* est déjà un pas vers une meilleure interprétation des résultats (Girondot and Guillon, 2018).

Enfin, lorsque de nombreux facteurs sont testés et combinés, et les mesures multipliées, le résultat, pour aussi probant qu'il puisse paraître, ne doit être considéré que comme un indice, et non pas comme la preuve qu'un effet existe. En effet, la démarche scientifique consiste à formuler une hypothèse à partir d'observations, avant de la tester avec de nouvelles observations. Une étude envisageant des effets aussi nombreux que ceux testés par la multiplication et la combinaison de facteurs et de mesures doit être considérée comme exploratoire (Forstmeier et al., 2017). Un tel travail est souvent indispensable pour arriver à déceler d'éventuels effets. Cependant, une nouvelle étude, centrée sur ces éventuels effets et utilisant de nouvelles données, reste nécessaire pour apporter des éléments probants en faveur d'une hypothèse quelle qu'elle soit.

---

[1] La puissance d'un test est sa capacité à détecter une différence si elle existe. La puissance dépend du nombre de sujets inclus et de la taille de l'effet relativement à la variance liée à l'échantillonnage.

[2] La notion de biais en statistique désigne une déviation systématique par rapport à la vraie valeur. Ici les résultats sont biaisés car un effet significatif est plus souvent trouvé que ce qui est attendu.

Conclusions

Nous avons vu dans un premier temps combien il était facile de trouver un résultat significatif alors que pourtant les données avaient été générées au hasard. Les raisons en sont multiples. En particulier la multiplication des tests conduit à détecter bien trop souvent des effets qui sont des faux positifs. Une première solution, consiste à corriger les seuils de significativité de manière à prendre en compte la multiplication des tests (Bonferroni, *False Discovery Rate*). Une seconde solution, mise en œuvre dans cet article, consiste à générer des jeux de données purement aléatoires mais ayant les mêmes caractéristiques que les données réelles et à leur appliquer les mêmes méthodes que celles utilisées dans l'étude pour évaluer le nombre de potentiels faux-positifs.

Il convient de noter que notre message s'applique à de nombreuses procédures statistiques, même lorsqu'elles paraissent très complexes et utilisent des corrections pour des cofacteurs. L'utilisation des méthodes semi-bayésiennes (Braun et al., 2014), des AIC ou AICc ou BIC (Burnham and Anderson, 2002; Girondot and Guillon, 2018) ou des *quartiles' mean score* (Braun et al., 2009) ne devrait pas empêcher de se pencher sur le problème des faux-positifs. Il faut lire de façon critique ce qui est publié et savoir détecter les mauvaises pratiques. Et surtout les éviter soi-même.

Contacts

Marc Girondot et Jean-Michel Guillon

marc.girondot@u-psud.fr, Jean-Michel.Guillon@u-psud.fr

Ecologie, Systématique, Evolution, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91400 Orsay, France

Bibliographie citée

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289-300.

Braun, J.M., Kalkbrenner, A.E., Just, A.C., Yolton, K., Calafat, A.M., Sjodin, A., Hauser, R., Webster, G.M., Chen, A., Lanphear, B.P., 2014. Gestational exposure to endocrine-disrupting chemicals and reciprocal social, repetitive, and stereotypic behaviors in 4- and 5-year-old children: the HOME study. Environmental Health Perspectives 122, 513-520.

Braun, J.M., Yolton, K., Dietrich, K.N., Hornung, R., Ye, X., Calafat, A.M., Lanphear, B.P., 2009. Prenatal bisphenol A exposure and early childhood behavior. Environmental Health Perspectives 117, 1945-1952.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach. Springer-Verlag, New York.

Burnham, K.P., Anderson, D.R., 2014. P values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology 95, 627-630.

Dunn, O.J., 1959. Confidence intervals for the means of dependent, normally distributed variables. Journal of the American Statistical Association 54, 613-621.

Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Natl Sci Rev 1, 293-314.

Fisher, R.A., 1934. Statistical Methods for Research Workers, 5th edition ed, London, UK.

Forstmeier, W., Wagenmakers, E.-J., Parker, J.H., 2017. Detecting and avoiding likely false-positive findings : a practical guide. Biological Reviews 92, 1941-1958.

Geiss, F., Bino, G.D., Blech, G., Nørager, O., Orthmann, E., Mosselmans, G., Powell, J., Roy, R., Smyrniotis, T., Town, W.G., 1992. The EINECS inventory of existing chemical substances on the EC market. Toxicological 37, 21-33.

Girondot, M., Guillon, J.-M., 2018. The w-value: An alternative to t- and χ^2 tests. Journal of Biostatistics & Biometrics 1, 1-4.

Guxens, M., Lubczynska, M.J., Muetzel, R.L., Dalmau-Bueno, A., Jaddoe, V.W.V., Hoek, G., van der Lugt, A., Verhulst, F.C., White, T., Brunekreef, B., Tiemeier, H., El Marroun, H., 2018. Air pollution exposure during fetal life, brain morphology, and cognitive function in school-age children. Biological Psychiatry. In press.

John, L.K., Loewenstein, G., Prelec, D., 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci 23, 524-532.

Yoccoz, N.G., 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72, 106-111.

.