Projet de thèse : contrats doctoraux SFR-Agorantic 2016 – 2019

Codirecteur de thèse : Philippe Martin (Prof. UAPV, UMR ESPACE)
Codirectrice de thèse : Delphine Blanke (Prof. UAPV, LMA)
Encadrant :

Correspondant : Philippe Martin (philippe.martin@univ-avignon.fr) : 04.90.16.26.95

Titre : ***Modélisations parétiennes à paramètres multiples de phénomènes hiérarchiques contraints mesurés à différentes résolutions. Recherche sur le sens de ces paramètres à partir d'exemples pris dans différentes disciplines.***

Mots clefs : Pareto, hiérarchie, fractale, résolution, écologie, géographie, statistique

Profil du candidat

Le (ou la) candidat(e) devra être titulaire d'un master II ou équivalent, être autonome, volontaire et capable de s'insérer dans une équipe de chercheurs et d'enseignants chercheurs — d'horizons variés — qui travaillent depuis quelques années sur ces questions, mais dans des domaines différents (géographie, statistique, mathématique, informatique, etc.). La thèse devra être réalisée en partie en relation avec les travaux de l'équipe de Statistique du LMA et ceux de l'UMR ESPACE.

Le candidat doit avoir de bonnes connaissances en statistique et en géographie théorique et quantitative. En particulier il/elle doit être capable de programmer différentes solutions logicielles nécessaires à la thèse dans la mesure où certains travaux le (ou la) conduiront à manipuler des bases de données importantes (big data). Par ailleurs il/elle doit posséder une connaissance minimale en écologie, géographie, sociologie, etc. même si, lors de la thèse, des apprentissages disciplinaires pourront être réalisés. Cette thèse a donc une dimension multidisciplinaire importante et s'inscrit dans l'approche interdisciplinaire de la SFR Agorantic.

Elle est conçue comme une recherche théorique sur la modélisation parétienne de hiérarchies contraintes. Laquelle sera appliquée à différentes séries de données issues de champs différents. De cette mise en œuvre, nous espérons une compréhension fine du sens des paramètres utilisés, dans chacune des situations disciplinaires envisagées.

Ce travail pourra se faire parallèlement avec la réalisation de mémoires ou de stages longs (supérieurs à deux mois).

Cadre institutionnel

Le doctorant sera basé à Avignon. Ce travail sera réalisé à l'UMR ESPACE 7300 et dans le cadre du laboratoire de Mathématiques de l'UAPV et de son équipe de Statistique. Il vient à l'appui de programmes en cours ou soumis de l'UMR ESPACE. En particulier il entre dans une réflexion sur la fractalité des phénomènes conduisant à des situations critiques (criticité des basses eaux – sécheresse).

Ce travail abordera aussi le caractère complexe de certains systèmes anthropiques comme le réseau urbain mondial (ANR Franco-Brésilienne AMERICApolis en évaluation finale) ou numériques comme des résultats d'enquêtes (pétitions, etc.), etc.

Contexte et enjeux

Wilfried Pareto est connu pour son cours d'économie tout autant que pour sa loi des 80 – 20 qui traduit, très didactiquement, l'idée d'une forte inégalité : ainsi 80 % des possédants disposent-ils de 20 % des richesses mondiales alors que 20 % des plus riches détiennent eux, 80 % des avoirs planétaires, mais 1 % des plus fortunés possède plus de 50 % des richesses, etc. C'est donc bien d'une hiérarchie dont rendent compte ces observations empiriques.

Ce type de hiérarchie se rencontre non seulement en économie (niveau de capital, revenus, PNB par habitant, etc.), mais aussi dans de très nombreux autres domaines, comme en particulier le fonctionnement du web (réseaux, flux, files d'attente, etc.) et en géographie (réseaux urbains, incendies de forêt, etc.).

Plus techniquement, les séries statistiques représentant ces fortes hiérarchies peuvent être traduites empiriquement dans un graphique bilogarithmique — qui croise une série expérimentale et une série théorique (le rang, une fréquence, etc.) — par un alignement de points, auquel peut être ajusté un modèle de puissance dont l'exposant, dit de Pareto, est le seul paramètre statistique. En géographie, ce type d'approche (sous une forme discrète) est connu sous le nom de loi de Zipf, laquelle a été réinterprétée par B. Mandelbrot comme étant l'expression d'une fractalité, d'une organisation en échelle.

Cependant, expérimentalement, cette modélisation avec un seul paramètre n'est pas toujours satisfaisante, et cela même si on envisage la loi de Pareto comme une loi L-stable. Très souvent, lorsque les catalogues de données sont importants et considérés dans leur intégralité, il apparaît ce qui est empiriquement une courbure et qui traduit fonctionnellement une atténuation de la hiérarchie par rapport à ce qu'elle aurait pu être si elle avait suivi intégralement une loi de Pareto. Ce que nous dénommerons : hiérarchie capée, ou « contrainte », mais par quoi ?

Un mauvais ajustement par la loi de Pareto peut conduire les statisticiens à envisager soit une rupture de modèle et donc ajuster seulement le modèle sur les données asymptotiques ou bien envisager un modèle plus complexe de type mélange. Ces démarches posent alors des problèmes de détection de la rupture, d'estimation des paramètres de mélange, et d'interprétation à donner aux différents modèles ajustés.

Il convient donc de reprendre la tentative qu'avait fait B. Mandelbrot de modéliser cette « courbure » à partir d'exemples littéraires (dénombrement de mots) et en s'appuyant sur la théorie de l'information de Shannon, afin :

-1 — de disposer d'une modélisation statistique robuste dont les paramètres (deux probablement) seront intrinsèquement compris et explicités ;

-2- d'accéder à une compréhension de ces structures ubiquistes qui ne soit pas contrainte par le type de substrats ou de milieux dans lesquels elles se manifestent.

Ces questions de construction formelle d'un modèle parétien statistique « contraint » et d'explicitation quel que soit le domaine des paramètres du modèle, restent encore aujourd'hui largement ouvertes malgré leur importance dans la compréhension de certains phénomènes ou fonctionnements. C'est en particulier le cas en géographie où l'usage de modèles très basiques (loi rang taille, etc.) rend très inopérant toute réflexion un peu fine, en raison de la faible adéquation de la modélisation aux séries.

Cela étant de nombreuses solutions partielles existent, mais sans que les raisons de ces multiples solutions spécifiques ne soient bien comprises. Il semble manquer une sorte « d'unification » qui dessinerait une cartographie des solutions en rapport avec des raisons objectives (liées aux séries statistiques) à mettre en œuvre ; rien n'interdisant au final d'essayer de simplifier ou de rationaliser ce qui existe à partir d'une compréhension nouvelle de l'ensemble.

Pour essayer d'aller plus loin, et pour pouvoir diffuser de meilleures pratiques, en particulier en SHS, dont on espère une bien meilleure compréhension des phénomènes, il convient aussi de mettre à disposition de thématiciens des outils adaptés de traitement de l'information, de modélisation statistique. Il s'agit donc aussi, par ce biais, de travailler au décloisonnement des disciplines.

Ces recherches ont été initiées au travers d'un financement Agorantic en 2015 (Projet Pareto). Elles se poursuivront, si le projet est retenu par l'ANR, dans l'ANR AMERICAPolis (piloté par F. Moriconi-Ebrard DR CNRS, professeur invité au Brésil) à laquelle participent les deux co-directeurs et des membres possibles du Comité de thèse.

Objectifs et méthode

Le premier objectif est donc de modéliser des hiérarchies capées avec un modèle limité à deux paramètres. Différents travaux réalisés par les co-directeurs semblent montrer que cela est tout à fait possible et que plusieurs séries statistiques numériquement importantes (plusieurs dizaines de milliers de valeurs) sont particulièrement bien décrites avec de telles solutions.

Le (ou la) candidat(e) devra donc s'approprier ces considérations, les mettre en œuvre tout en les développant. Ceci nécessitera de construire des outils informatiques susceptibles de traiter des masses importantes de données.

Techniquement, il sera proposé de travailler d'abord sur des données relevant du champ de la géographie physique, puis sur des données relevant du champ de la géographie humaine (en particulier des données de recensements de populations), puis sur des données issues du fonctionnement de la société numérique.

Cet ordre s'explique par le caractère vraisemblablement plus assuré des données physiques, car pensées comme dépendant d'un processus physique peu sujet à différents biais. De plus les données physiques, dans la mesure où elles ont été acquises à haute résolution peuvent être transformées en données à d'autres résolutions plus grossières, ce qui devrait permettre de voir comment ces lois statistiques (et donc leurs paramètres) se transforment en fonction de l'échelle d'appréciation du phénomène. La compréhension des phénomènes de transformation des lois en fonction de l'échelle de la résolution de l'information pourrait être un préalable à une certaine unification du champ.

Ces éléments seront utiles ensuite pour aborder d'autres données en particulier issues de la société numérique et plus largement des SHS qui portent souvent la trace de choix explicites ou implicites dont les conséquences sur les modèles ajustés sont généralement inconnues. Des tests pourront être envisagés pour détecter les manques éventuels de robustesse dans ces méthodes. Tests pour connaitre la robustesse des déterminations numériques des paramètres de ces modèles. Mais aussi tests pour aborder les conséquences sur les paramètres de choix dans la construction des données, parfois fort peu variables.

Il s'agira de voir, dans un second temps, s'il est possible d'aller (ou pas) vers une solution générique et d'essayer de comprendre l'adéquation du modèle à une série statistique, donc le sens statistique et thématique des paramètres mis en œuvre. Ce n'est qu'à ce niveau, semble-t-il, que la diversité des séries statistiques permettra de comprendre la multiplicité des modèles et que la multiplicité des ajustements permettra d'aborder le sens des paramètres utilisés (par exemple en les rapprochant d'autres considérations) ; cela représenterait une avancée très importante pour une discipline comme la géographie et plus largement pour la compréhension de problèmes issus de la société numérique et des SHS.

Cette approche multidisciplinaire, ou du moins touchant à différents domaines, est aujourd'hui possible dans la mesure où la société numérique met à disposition des masses absolument

considérables de données pour lesquelles on ne dispose pas toujours des bons outils de traitement. Cette thèse aura ainsi aussi pour objectif de concevoir et de produire de tels outils.

Ceci conduira à traiter les problèmes suivants :

- Est-il possible de modéliser, si ce n'est toutes les hiérarchies, mais du moins une grande partie d'entre elles et en particulier les hiérarchies capées, avec un modèle d'essence parétienne à deux paramètres ?
- À partir d'une utilisation multidisciplinaire de cette modélisation qui mettrait en œuvre plusieurs variantes de modèles parétiens est-il possible d'établir des relations univoques entre variantes et types de séries ? Les caractéristiques des séries permettant peut-être de comprendre la multiplicité des variantes ?
- Quel sens donner aux paramètres utilisés, peut-être dans le cadre de la fractalité, à partir d'une utilisation multidisciplinaire de ces modèles ? Peut-on les estimer de manière plus efficace qu'avec une méthode d'ajustement linéaire classique ?
- Renvoient-ils chacun, comme l'exposant de Pareto à un concept spécifique (celui de hiérarchie) —on pourrait penser par exemple à une dimension fractale variable dans l'ordre des échelles, et dans cette hypothèse au besoin d'un travail plus théorique d'articulation et d'approfondissement de façon à unifier certains aspects statistiques et fractals — ou ne sont-ils qu'empiriques et donc spécifiques aux distributions ?

Verrous

- Le développement et la validation d'une modélisation d'essence parétienne à deux paramètres valide pour de nombreuses ou pour toutes (?) les situations ;
- La recherche de solutions techniques pour déterminer des valeurs robustes des paramètres ;
- La réalisation d'outils de traitement ergonomiques ;
- La recherche d'une solution générique sous une forme diffusable et utilisable par des non-spécialistes.
- L'intégration de ces approches dans le cadre de la géométrie fractale.

Retombées attendues

- Des outils relativement simples de modélisation utilisables dans le secteur de Sciences Humaines et Sociales et plus largement pour aborder des problèmes de la société numérique.
- Une meilleure compréhension de ces phénomènes de hiérarchies capées ou de hiérarchies « contraintes » qui semblent très fréquents ou généraux ;
- Des explications théoriques et thématiques sur ces « contraintes », sur les causes de ces courbures et de leur variation ;
- Une compréhension de la variation des paramètres (donc de la forme des distributions) dans le temps, dans l'étendue (surface terrestre) et dans les échelles en fonction de la résolution de l'information ;
- La formalisation des phénomènes de transformation des lois en fonction de l'échelle de résolution de l'information ;
- L'intégration de ces réflexions dans un cadre plus large peut être d'essence fractale.

Éléments bibliographiques

Adler R. J., Feldman R. E., Taqqu M. S. (Dir), 1998, *A pratical guide to heavy tails. Statistical techniques and applications*. Birkhäuser, Boston, 533 p.

Blanke, D. 2015, Notes sur la loi de Pareto et ses variantes possibles, document de travail, 11 pages.

Forriez M., Martin Ph., 2009, Structures hiérarchiques en géographie : des modèles linéaires aux modèles non linéaires (lois de puissance et corrections log-périodiques), in Foltête, Jean-Christophe (s.d.), Huitièmes Rencontres de ThéoQuant, Besançon, THEMA, 10 p. http://thema.univ-fcomte.fr/theoq/pdf/2007/TQ2007%20ARTICLE%2051.pdf

Feuerverger, A, Hall, P., Estimating a tail exponent by modelling departure from a Pareto distribution, Annals of Statistics 1999, Vol. 27, No. 2, p. 760-781

Mahanti A, et al., 2013, A tale of the tails: Power-laws in internet measurements, Network, IEEE, Vol.27, Issue 1, p.59-64.

Martin Ph., 2013, Statistique dynamique d'une série parétienne d'observations approche méthodologique. 11e rencontre de ThéoQuant, Besançon, livre des résumés, http://thema.univ-fcomte.fr/theoq/pdf/resumes/TQ2013%20RESUMES.pdf p.209-210.

Martin Ph., Redjimi M., 2014, Évolution du réseau de peuplement algérien depuis l'Indépendance Approches parétienne et fractale parabolique. *Aux frontières de l'urbain : Petites villes du monde*, Collection : Actes Avignon — ISBN : 978-2-9105-4509-1, p.374-395.

Martin Ph., (2016) – Modélisation des longueurs des périodes sans pluies supérieures à différents seuils de la chronique de Marseille (1864-2008), *Physio-Géo*, Volume 10, 1, 81-104.

Martin Ph., (en préparation) — Modélisation parétienne des longueurs des périodes sans pluies supérieures à un seuil choisi sous climat méditerranéen (Marseille : 1864 – 2008).

Mandelbrot B., 1997, *Fractales, hasard et finance*. Coll. : Champ, Flammarion éditeur, Paris, 246 p.

Newman, M.E.J., 2006, Power laws, Pareto distributions and Zipf's law, preprint arXiv 0412004, http://arxiv.org/abs/cond-mat/0412004v3, 28 pages

Pumain D., 2006, (Dir), *Hierarchy in natural and social sciences.* Methodos séries, vol.3, Springer éditeur, Berlin, 243 p.

Saichev A., Malevergne Y, Sornette D., 2010, *Theory of Zipf's law and beyond*. Springer Verlag, Berlin, 169 p.

Zajdwenweber D., 2009, *Économie des extrêmes. Krachs, catastrophe et inégalité*. Coll. Champs essais, Flammarion, Paris, 236 p.

Zipf G.K., 1949, *Human behaviour and principle of least effort*, Cambridge, MA: Addison-Wesley.