Enseigner les outils de modélisation statistiques des territoires dans un contexte international et pluridisciplinaire : enjeux et dilemmes du choix des exemples d'application

Gbetoton Nadège Djossou, Claude Grasland

Enseigner les outils de modélisation statistiques des territoires dans un contexte international et pluridisciplinaire : enjeux et dilemmes du choix des exemples d'application

AUTEUR·ES
Gbetoton Nadège DJOSSOU,
Claude GRASLAND

RÉSUMÉ
Les statistiques constituent une composante essentielle du bouquet d'outils utilisé par les sciences territoriales. Mais leur enseignement est souvent difficile à assimiler pour les débutants en raison du choix d'exemples inadaptés aussi bien pour la partie théorique (cours) que pour les applications (travaux dirigés). Nous proposons de discuter les enjeux pédagogiques mais aussi politiques et symboliques du choix des exemples à travers le cas d'une initiation aux méthodes d'analyse de variance et de régression linéaire multiple destinée à des enseignants-chercheurs de huit pays (Bénin, Burkina Faso, Côte d'Ivoire, France, Mali, Niger, Sénégal, Togo) et de quatre disciplines (géographie, démographie, économie, sociologie). Quatre dilemmes relatifs au choix des exemples sont discutés qui concernent respectivement (i) le choix d'exemples simples ou complexes, (ii) le choix d'exemples concrets ou abstraits, (iii) le choix de données spatialement agrégées ou individuelles et enfin (iv) le choix des terrains d'études situés au nord ou au sud.

MOTS CLÉS
sciences du territoire, statistique, pédagogie, régression multiple, Afrique, France, R

ABSTRACT
Statistics constitute an essential component of the toolbox used in territorial sciences. But teaching them is often difficult for beginners to assimilate because of the choice of inappropriate examples, both for the theoretical part (courses) and for the applications (tutorials). We propose to discuss the pedagogical as well as the political and symbolic issues involved in the choice of examples through the case of an introduction to the methods of variance analysis and multiple regression intended for teacher-researchers from eight countries (Benin, Burkina Faso, Côte d'Ivoire, France, Mali, Niger, Senegal, Togo) and four disciplines (geography, demography, economics, sociology). Four dilemmas relating to the choice of examples are discussed which concern respectively (i) the choice of simple or complex examples, (ii) the choice of concrete or abstract examples, (iii) the choice of spatially aggregated or individual data, and finally (iv) the choice of study sites located in the North or the South.

KEYWORDS
Territorial sciences, Statistics, Pedagogy, Multiple regression, Africa, France, R

CONTEXTE : L'ÉCOLE D'ÉTÉ DU CIST EE·2022-2023

L'école d'été « Outils et méthodes des sciences territoriales », qui a été conçue par le CIST, l'IRD, l'INED et l'Université de Parakou, rassemble trente enseignants-chercheurs et ingénieurs de huit pays (Bénin, Burkina Faso, Côte d'Ivoire, France, Mali, Niger, Sénégal, Togo) et de quatre disciplines (géographie, démographie, économie, sociologie) pour élaborer conjointement un ensemble de cours et travaux dirigés selon une procédure d'échange et de coconstruction.

Au cours d'une première session de travail à Paris en juin 2022, les formateurs, qui avaient bénéficié d'une formation initiale au logiciel *R*, se sont mis d'accord sur le choix d'une quinzaine de modules couvrant les domaines de la statistique, de la cartographie et de l'analyse spatiale. Pour chaque module, un plan provisoire du contenu a été défini, les équipes ont précisé les logiciels qui seraient utilisés (*R* par défaut, sinon *QGIS*, *Magrit*, etc.) et réfléchi aux jeux de données qui pourraient être mobilisés pour illustrer le cours et construire des travaux dirigés. Deux coresponsables ont été ensuite attribué à chaque module, un issu de France et l'autre d'Afrique, pour préparer leur contenu, dans l'objectif de les tester lors de la seconde phase de l'école d'été à Ouidah (Bénin), qui a duré deux semaines en février-mars 2023. Des stagiaires (doctorants) se sont ajoutés à l'équipe des enseignants-chercheurs / ingénieurs lors de la deuxième semaine de l'école d'été de Ouidah (Bénin) et les modules ont pu être testés avec eux.

Après des débats animés sur le degré d'approfondissement des modules que l'on pourrait assimiler en deux semaines (« beaucoup de peu » ou « peu de beaucoup »), les organisateurs sont arrivés à un compromis consistant à centrer la première semaine sur les modules fondamentaux et la seconde sur les modules plus avancés qui ne font l'objet que d'une simple initiation (fig. 1). Parmi les modules avancés pour lesquelles l'objectif n'est pas de fournir une connaissance complète mais une simple initiation, figurent la régression multiple et l'analyse de variance (module MOD1) dont nous étions responsables et que nous prenons en exemple ici. Il visait à modéliser une variable dépendante Y quantitative par des variables indépendantes de type quantitatif (X1, …, Xk) ou qualitatif (Q1, …, Qn). Ce module ne disposait en pratique que de 2h de cours et 2h d'application sur *R*, ce qui est évidemment trop peu pour une formation aboutie mais *a priori* suffisant pour donner aux stagiaires une idée générale des objectifs de la méthode et de ses possibilités. Restait à choisir les bons exemples d'application, ce qui a suscité plusieurs dilemmes, à

savoir : (i) le choix entre deux modèles pédagogiques alternatifs (le choix d'exemples simples / complexes ou le choix d'exemples concrets / abstraits), (ii) le choix de données (données spatialement agrégées ou individuelles ; terrains d'études situés au nord ou au sud).

Figure 1. Programme de l'école d'été du CIST EE·2023 (Ouidah, Bénin)

Semaine 1	Lundi	Mardi	Mercredi	Jeudi	Vendredi
8h30-10h30	DON 1-2-3 Types de données	EXP1 : Univarié	EXP2 : Bivarié	SPA1 : Semis de Points	EXP3 : Multivariée
11-13		CART1-2 Carto thématique	EXP2 : Bivarié	SPA2 : Distance & Accesibilité	
14-16	R ou QGIS	R ou QGIS	R ou QGIS	R ou QGIS	R ou QGIS
16-18	DON- Application (parallèles)	CART1-2 Application	EXP1-2 Application	SPA1 & SPA2 Application	EXP3 Application

Semaine 2	Lundi	Mardi	Mercredi	Jeudi	Vendredi
8h30-10h30	Ateliers de terrain	Ateliers de terrain	MOD1 Y Quantitative	MOD2 Y Qualitative + Multiniveau	Bilan et perspective
11-13	ou Cours Stat & R	ou Cours Stat & R	CART3 Carto Dynamique	CART4 Carto sous R	
14-16	Ateliers de terrain	Ateliers de terrain	MOD1 Applications	MOD2 Applications	
16h30-18	ou Cours Stat & R	ou Cours Stat & R	CART3 Applications	CART4 Applications	

DON :	Types de données	SPA :	Analyse spatiale
EXP:	Statistiques exploratoires	CART :	Cartographie
MOD :	Modélisation statistique		

LE CHOIX ENTRE DEUX MODÈLES PÉDAGOGIQUES ALTERNATIFS

Le module dont nous étions responsables (MOD1) avait pour objectif d'effectuer le pont entre les approches de statistiques descriptives (EXP1, EXP2, EXP3) et les approches explicatives fondées sur la famille des modèles de régression linéaire multiple (MOD1), de régression logistique (MOD2) et de modélisation multiniveau (MOD3). Il fallait donc s'assurer au préalable que les stagiaires maîtrisaient bien les tests de significativité de la relation entre deux variables quantitatives (Pearson et Spearman) ou entre une variable quantitative et une variable qualitative (test d'égalité des moyennes de Student, analyse de variance et test de Fisher). Les rappels relatifs à la relation entre deux variables qualitatives (Chi-2) figuraient quant à eux dans les modules suivants (MOD2 et MOD3).

Au vu du faible volume horaire dont nous disposions, nous ne pouvions pas développer de multiples exemples ; il était essentiel d'en trouver qui combinent à la fois de bonnes propriétés théoriques (e.g. présence de valeurs exceptionnelles à retirer ou d'hétéroscédasticité à corriger par une transformation log-linéaire) et un intérêt empirique des résultats (afin de maintenir l'attention et de faciliter la mémorisation).

L'approche pédagogique s'est alors confrontée au dilemme de choisir entre deux méthodes pédagogiques proposées par deux acteurs qui ont opté pour des solutions diamétralement opposées en ce qui concerne l'initiation aux méthodes de modélisation statistique d'une variable quantitative : le psychologue Denis (2020), qui a opté pour des méthodes d'initiation basées sur des données fictives et construites sur mesure, et le géographe (Taylor, 1977), qui a quant à lui opté pour des données concrètes. Les deux auteurs constituent d'excellents pédagogues et d'excellents statisticiens mais lequel fallait-il suivre pour construire un cours efficace et intéressant dans le contexte de l'école d'été du CIST EE·2023 ?

Le choix de ces deux modèles pédagogiques a pour avantage de permettre à des débutants de disciplines diverses d'appliquer la théorie à deux disciplines différentes ainsi qu'à différents types de données.

Daniel J. Denis : des exemples psychologiques fictifs construits sur mesure

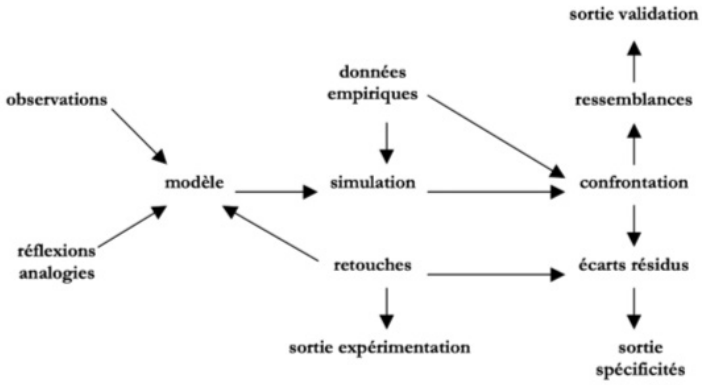
Pour l'auteur du remarquable ouvrage *Univariate, Bivariate, And Multivariate Statistics Using R: Quantitative Tools For Data Analysis And Data Science* (Denis, 2021), il ne fait aucun doute que les meilleurs exemples pédagogiques pour découvrir une méthode statistique sont ceux que l'on a inventés soi-même afin d'y introduire précisément le niveau de difficulté souhaité mais pas davantage. S'agissant d'un ouvrage de statistique adossé sur l'apprentissage d'un logiciel (*R* ou *Python*), les données servant d'exemple sont le plus souvent des vecteurs d'une vingtaine d'individus qu'il est simple de saisir au clavier pour tester ensuite les différentes formes de modélisation. Il est facile, à partir de là, d'introduire des cas théoriques importants tels que la présence d'une valeur exceptionnelle, d'une distribution asymétrique ou bimodale, etc. Et, plus important encore, de montrer des exemples précis de multicolinéarité entre variables explicatives pouvant aboutir à ce qu'une variable initialement significative ne le soit plus lorsqu'on en ajoute une nouvelle ou que, réciproquement, une variable non significative de façon isolée le devienne avec l'ajout d'une nouvelle. Bref, les cas complexes d'interaction entre variables explicatives qui constituent le cœur de la régression multiple sont plus simples à expliquer lorsque l'on crée soi-même un exemple parfaitement illustratif.

Pour autant, et c'est l'une de ses grandes qualités, Denis (*ibid*.) prend toujours soin de donner une signification aux variables fictives qu'il a créées, pour en faciliter l'apprentissage. Les exemples pris dans le domaine de la psychologie cognitive (e.g. compétence acquise selon l'enseignant et le choix du manuel) sont d'ailleurs souvent repris d'un chapitre à l'autre ce qui permet de les assimiler sur les cas simples (statistique univariée ou bivariée) avant de les retrouver sur des cas plus complexes (analyse multivariée, régression multiple). Au total, la solution proposée consiste certes à retenir des exemples fictifs mais en leur donnant une signification très concrète. Avec une particularité importante : il s'agit d'exemples relatifs à des individus (personnes) et non à des agrégats, de sorte que l'hypothèse de causalité reste probable même si elle n'est pas forcément démontrable.

Peter J. Taylor : un exemple géographique concret mobilisant des causes déterministes

Le géographe Peter J. Taylor, connu actuellement pour ses travaux sur l'établissement d'une liste des villes mondiales dans le cadre du projet Globalization and World Cities[1], a été au début de sa carrière un formidable pédagogue de l'usage des méthodes quantitatives en géographie pour lesquelles il a rédigé un manuel également remarquable (1977). C'est toutefois dans une publication pédagogique moins connue de 1980 que l'on trouve un article plus centré sur les problèmes de pédagogie des modèles statistiques intitulé « *A Pedagogical Application of Multiple Regression* » (1980) qui a été ultérieurement repris par l'un des auteurs, tant l'exemple proposé présente des qualités remarquables, au moins pour les géographes (Grasland, 1995). La démonstration suit en effet une parfaite démarche hypothético déductive où chaque introduction d'une nouvelle variable est suivie d'une analyse des résidus et d'une retouche du modèle initial (fig. 2) suivant les préceptes d'un modèle de la modélisation formulés par Durand-Dastès (1991).

Figure 2. Un modèle de la modélisation (Durand-Dastès, 1991)



La démonstration est presque trop belle pour être vraie tant elle permet de couvrir la plupart des difficultés théoriques rencontrées en statistique sur ce chapitre difficile de la régression multiple. La variable dépendante est le volume moyen annuel de précipitations de trente stations météorologiques de Californie dans les années 1950-1975. Après avoir modélisé (1) l'effet de trois variables quantitatives ayant des effets déterministes sur celle-ci (altitude, latitude, distance à la mer), l'auteur introduit (2) une variable qualitative (situation d'abri par rapport aux vents dominants) qui va fortement augmenter le pouvoir explicatif du modèle tout en rendant non significative l'une des variables initiales (altitude). Puis l'élimination de deux stations (3) présentant des résidus exceptionnels augmente encore le coefficient de détermination tout en rétablissant le pouvoir explicatif de la variable explicative éliminée (fig. 3).

Figure 3. Modélisation des précipitations en Californie (d'après Taylor, 1980 ; Grasland, 1995)

	Dependent variable:		
	Précipitations (en mm)		
	(1)	(2)	(3)
Latitude (degrés N)	87.893***	87.883***	77.836***
	(20.175)	(16.682)	(10.155)
Altitude (m)	0.339***	0.183*	0.275***
	(0.101)	(0.094)	(0.052)
Distance à la mer (km)	-2.265***	-0.852	-0.899**
	(0.577)	(0.617)	(0.330)
Abri (Oui/Non)		-401.351***	-284.514***
		(111.186)	(60.871)
Constant	-2,609.336***	-2,493.660***	-2,229.328***
	(741.251)	(613.735)	(369.080)
Observations	30	30	28
R^2	0.600	0.737	0.887
Adjusted R^2	0.554	0.695	0.867
Residual Std. Error	281.769 (df = 26)	232.978 (df = 25)	123.545 (df = 23)
F Statistic	12.997*** (df = 3; 26)	17.516*** (df = 4; 25)	44.959*** (df = 4; 23)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

Ce qui rend cet exemple particulièrement efficace sur le plan pédagogique est le fait d'introduire des causalités intuitives qui s'enchaînent et dont les résidus sont visualisés cartographiquement et permettent de choisir les variables ultérieures à ajouter.

1 [lboro.ac.uk/microsites/geography/gawc].

CHOIX DES TERRAINS ET DES NIVEAUX D'AGRÉGATION DES EXEMPLES RETENUS

S'il nous est apparu intéressant de retenir les deux approches précédentes pour dispenser le cours de 2h, nous avons souhaité proposer des applications qui soient plus en phase avec les attentes des formateurs en termes de terrains (de préférence africains) et en termes disciplinaires. Sur ce dernier point, il nous est apparu important de proposer deux exemples de niveaux différents d'agrégation, l'un portant sur des données individuelles et l'autre sur des données spatiales agrégées. La communication proposée au colloque du CIST porte plus précisément sur ces deux exemples et leur réception auprès du public. On indique juste ici les jeux de données mobilisées et la problématique adoptée.

Modélisation du développement des pays africains en 2017-2018

Le premier exemple du module a utilisé des données décrivant les pays africains en 2017-2018 à l'aide d'indicateurs tirés du rapport mondial sur le développement humain de 2020[2] complété par quelques variables tirées de la base des pays du Monde du CEPII[3]. La base de données possède plusieurs avantages :

– elle a déjà été utilisée lors de l'initiation à *R* dispensée aux formateurs de l'EE au printemps 2022 pour les premiers cours portant sur l'analyse de variance ou la régression linéaire simple ;

– elle est de taille assez réduite (49 pays) mais suffisante pour mener des analyses statistiques multivariées : on peut faire des analyses de la variance en utilisant par exemple la variable « découpage Afrique en 5 régions » ou la variable booléenne « pays enclavé ou non » ;

– elle est assez complète mais contient quelques valeurs manquantes (notamment pour l'Érythrée ou le Sud Soudan) ce qui permet des exercices de prédiction issue d'un modèle de régression simple ou multiple ;

– elle comporte des pays de taille très différente ce qui pose la question de la pondération des analyses ;

– elle contient des valeurs exceptionnelles ou aberrantes et affiche des distributions non gaussiennes pour plusieurs indicateurs ce qui permet de poser des problèmes de transformation d'indicateurs ;

– elle peut facilement être mise à jour dans la mesure où le rapport sur le développement humain reprend chaque année la plupart des indicateurs antérieurs ;

– on peut ultérieurement construire des données diachroniques afin de suivre des évolutions.

La base de données a permis de faire la modélisation de la mortalité infantile des pays africains en fonction de leur produit intérieur brut (PIB), de leur taux d'urbanisation et de la transition géographique. Les stagiaires ont pu disposer à l'issue des travaux pratiques d'un programme typique d'analyse de régression sous *R* qu'ils pourront adapter à leurs propres données par la suite[4].

Modélisation des revenus des ménages du Bénin en 2018

Le second exemple de ce module de cours prévoyait d'utiliser des données de l'enquête de consommation des ménages (FinsScope) du Bénin en 2018 pour modéliser les revenus des ménages par tête en fonction de l'accès au crédit, du secteur d'activité du ménage, du sexe et de l'âge du chef de ménage, du milieu de résidence du ménage et de la situation de handicap ou non du chef de ménage. Les objectifs de cette enquête étaient de comprendre la population adulte du Bénin en termes de : moyens de subsistance et manière de gérer les revenus ; besoins et demandes financiers ; perceptions, attitudes et comportements financiers ; répartitions démographiques et géographiques ; et niveaux actuels d'accès et d'utilisation des services et produits financiers (Djossou *et al*., 2020).

Nous avons toutefois renoncé à l'utiliser lors de l'école d'été pour deux raisons. Premièrement, le temps imparti pour l'application était trop court pour introduire un second exemple qui aurait risqué de brouiller les acquis. Deuxièmement, la variable « revenu des ménages » était tantôt déclarée sous forme de valeur exacte, tantôt sous forme d'intervalle ce qui rendait son interprétation délicate. Il ne semblait pas judicieux de l'utiliser sans une présentation détaillée des problèmes d'analyse qu'elle soulevait. Nous avons toutefois laissé l'exemple en ligne sur le site web de l'école d'été[5] pour que les stagiaires les plus avancés puissent se l'approprier en s'appuyant sur la lecture d'ouvrage d'économétrie (Bourbonnais, 2018). Nous avons ajouté un second exemple de données individuelles en modélisant le nombre d'équipement des ménages du Bénin en 2013 en fonction de caractéristiques sociales et spatiales[6].

CONCLUSION : FAUT-IL « AFRICANISER » LES EXEMPLES ?

Une question difficile et certainement polémique s'est posée à l'ensemble des formateurs de l'EE·2022-2023 concernant le choix des données utilisées dans les cours et leurs applications. Fallait-il retenir uniquement des exemples pris dans les sept pays africains ? Fallait-il y ajouter des exemples français ? Fallait-il simplement prendre les exemples les plus efficaces sur le plan pédagogique, qu'ils soient ou non situés dans les pays des formateurs ? Chaque option peut en effet faire l'objet de critiques de nature différente :

– *ne retenir que les pays africains* peut donner l'impression que seuls les formateurs issus de ces derniers sont des « apprenants » ;

– *ajouter des exemples pris en France* peut donner l'impression d'un transfert de connaissance de type néocolonial ;

– *choisir des exemples sur la base de leur seule vertu pédagogique* peut conduire à reproduire à l'identique des exemples nés

2 [undp.org/fr/algeria/publications/rapport-sur-le-d%C3%A9veloppement-humain-2020].

3 [cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=6].

4 [ee-cist.github.io/MOD1_Yquanti/MOD1_Yquanti_exo1.html].

5 [ee-cist.github.io/MOD1_Yquanti/MOD1_Yquanti_exo3.html].

6 [ee-cist.github.io/MOD1_Yquanti/MOD1_Yquanti_exo4.html].

dans les pays du Nord à l'instar des fameux pourboires du garçon de café de New York que l'on retrouve dans tous les manuels de statistiques (Bryant & Smith, 1994 ; Cook & Swayne, 2007).

RÉFÉRENCES

Bourbonnais R., 2018, *Économétrie. Cours et exercices corrigés*, Malakoff, Dunod [10e éd.].

Bryant P. G., Smith M. A., 1994, *Practical Data Analysis: Case Studies in Business Statistics*, New York (NY), McGraw-Hill Professional.

Cook D., Swayne D. F., 2007, *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*, New York (NY), Springer.

Denis D. J., 2020, *Univariate, Bivariate, and Multivariate Statistics Using R: Quantitative Tools for Data Analysis and Data Science*, Hoboken (NJ), John Wiley & Sons.

Djossou G. N., Jacob N., Atchade Touwédé B., Abdelkrim A., 2020, *The Role of Formal, Informal, and Family Credit in the Business Performance of Young Entrepreneurs in Bénin*, Nairobi, PEP, « Working Paper » 2020-16 [portal.pep-net.org/document/download/34669].

Durand-Dastès F., 1991, « Le particulier et le général en géographie », *6e colloque de didactique de l'histoire, de la géographie et des sciences sociales (vol. 6)*, Lyon, INRP, p. 209-219.

Grasland C., 1995, « Modélisation et commentaire de documents. Application à l'étude des précipitations en Californie et des migrations entre les villes de plus de 50 000 habitants en France », *Feuilles de géographie*, no 16 [feuilles-de-geographie.parisnanterre.fr/1995/06/02/modelisation-et-commentaire-de-documents-application-a-letude-des-precipitations-en-californie-et-des-migrations-entre-les-villes-de-plus-de-50-000-habitants-en-france].

Taylor P. J., 1977, *Quantitative Methods in Geography: An Introduction to Spatial Analysis*, Boston (MA), Houghton Mifflin.

Taylor P.J., 1980, « A Pedagogic Application of Multiple Regression Analysis: Precipitation in California », *Geography*, 65(3), p. 203-212 [jstor.org/stable/40569273].

LES AUTEUR·ES

Gbetoton Nadège Djossou
Université de Parakou (Bénin)
nadeged2001@yahoo.fr

Claude Grasland
UPC – Géographie-cités / CIST
claude.grasland@parisgeo.cnrs.fr