

---

# La Médiennne : un compromis robuste entre la Moyenne et la Médiane

**Didier Josselin\* et Dominique Ladiray\*\***

\* *THEMA, UMR 6049, CNRS, Université de Franche-Comté, France, [didier.josselin@univ-fcomte.fr](mailto:didier.josselin@univ-fcomte.fr)*

\*\* *EUROSTAT, Unit 46, Statistical Indicators for Euro-zone Business Cycle Analysis, Luxembourg, [dominique.ladiray@cec.eu.int](mailto:dominique.ladiray@cec.eu.int)*

---

*RÉSUMÉ. Dans cet article, nous décrivons une nouvelle norme, combinaison linéaire des normes L1 et L2. Nous présentons la démarche qui nous a amené à la définir. Nous définissons la robustesse, discutons celle de la Moyenne et de la Médiane. Nous justifions la Médiennne, détaillons son calcul basé sur les variances estimées des deux valeurs centrales. Deux versions de Médiennne sont proposées, dont celle de Laplace qui avait montré son intérêt dès 1818. Nous comparons les robustesses des trois statistiques et appliquons la Médiennne à l'analyse de la pullulation du Campagnol Terrestre dans le Doubs.*

*MOTS-CLÉS. Médiennne, Moyenne, Médiane, Robustesse, Bootstrap, Lissage et filtrage spatial robuste.*

---

## 1. Pourquoi la Médiennne ?

La Médiennne est le fruit d'une réflexion à la confluence de deux disciplines : la Géographie et la Statistique, l'objectif commun étant d'améliorer l'efficacité et la pertinence des méthodes de filtrage des données, dans le domaine plus spécifique de l'analyse spatiale [FOT 2000].

Lorsque l'on souhaite analyser, à l'aide d'une méthode automatique numérique, une image brute ou une partition spatiale, le choix d'une méthode pertinente est crucial. Parmi les très nombreuses méthodes de « filtrage numérique » [COC 1995, GIR 1999], les filtrages spatiaux basés sur le voisinage, tels que les filtres passent-bas, sont largement utilisés. Suivant la fonction choisie, ils lissent l'image et réduisent son « bruit » (cas de la moyenne, par exemple) ou la filtrent en éliminant les « outliers » (cas de la médiane, par exemple). Un filtre moyen permet ainsi de souligner les zones géographiques hétérogènes, un filtre médian de révéler les zones homogènes et leurs contours. Lors d'une analyse locale de voisinage, les individus traités autour d'une cellule sont généralement peu nombreux. Les filtres spatiaux sont donc régulièrement confrontés à des distributions statistiques de faible effectif. Une petite modification de valeur ou une erreur de mesure, la présence de quelques individus éloignés du paquet de données, peuvent mettre à mal l'efficacité du filtre utilisé.

Les phénomènes observés en géographie montrent souvent une forte autocorrélation spatiale globale [CLI 1973] et/ou locale [ANS 1995], positive dans le cas où des individus proches se ressemblent, négative quand ils sont différents. Il est donc probable que le filtre rencontre des distributions avec des modes très marqués, assortis parfois d'« outliers ». Mais il est des cas où la distribution est moins typique. En cas d'absence d'autocorrélation spatiale, elle peut être quasi-uniforme ou multimodale. Si le filtre se trouve à la frontière de deux zones homogènes, la distribution risque d'être bimodale (et en plus asymétrique). Rechercher la position exacte d'une valeur centrale dans ces conditions peut poser problème, à cause de la relativité de sa signification réelle (sa valeur se situant loin des paquets d'individus) et de la forme des distributions observées (souvent éloignées des lois connues).

Faibles effectifs et formes imprévues des distributions statistiques justifient pleinement la recherche de méthodes de filtrage spatial robustes [JOS 2000].

## 2. Robustesse des estimateurs de centralité

La robustesse en statistique est un champ de recherche largement investi [HUB 1981, HAM, 1986, LEC 1987]. Un estimateur statistique est dit « robuste » s'il est peu sensible à un écart aux hypothèses statistiques de la loi. Il est dit « résistant » s'il est peu affecté par un petit nombre d'erreurs importantes ou par un plus grand nombre de petites erreurs [HOA 1983]. Les deux critères sont toutefois très liés. La Médienne repose sur le critère de résistance.

Les propriétés de la Moyenne sont bien connues : elle est sensible aux outliers ou à de fortes asymétries, notamment pour des distributions statistiques unimodales [WIL 2001]. On lui préfère, pour ce type de configuration, la Médiane, plus résistante. Généralement, le choix de la valeur centrale utilisée par les praticiens s'arrête à de telles considérations. Il existe pourtant de nombreux estimateurs robustes qui ont fait leurs preuves [AND 1972]. Ils sont malheureusement rarement utilisés dans la pratique.

Les cas de faible résistance de la Médiane existent, même s'ils sont rares. Ils correspondent souvent à des cas où la Moyenne est plus résistante. Le cas le plus évident est celui d'une distribution bimodale (discontinuités nettes entre deux zones homogènes). Plus les modes sont disjoints, plus l'influence de quelques individus situés entre les deux sera prégnante sur la localisation de la Médiane. La Moyenne, quant à elle, sera peu sensible à ces individus, à cause du poids important des deux modes dans son calcul. Entre ces deux cas extrêmes existe une multitude de distributions où les deux estimateurs sont plus ou moins robustes.

On voit bien la complémentarité des deux normes  $L1$  et  $L2$  pour un objectif de filtrage résistant. Deux questions se posent alors pour atteindre cet objectif. Comment incorporer les deux normes dans une même norme ? Comment estimer la résistance des deux valeurs centrales pour les combiner de manière efficace ? Nous proposons une solution simple de combinaison linéaire des deux indicateurs, pondérées par leurs résistances respectives. L'objectif est que la Médienne tende vers la Médiane lorsque celle-ci est plus résistante que la Moyenne, vers la Moyenne dans le cas inverse. Le critère de résistance que l'on considère est l'inverse de la variance estimée par un bootstrap.

## 3. Formalisation de la Médienne

La Médienne est donc un compromis entre la Moyenne et la Médiane. C'est une nouvelle norme combinant les normes  $L1$  et  $L2$ , fonctionnelle car la Médiane et la Moyenne mesurent le milieu d'une série de données, facilement applicable grâce aux méthodes informatiques de simulation et au bootstrap.

Soit  $\{x_1, x_2, \dots, x_n\}$  un échantillon de données,  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$  ce même échantillon ordonné.

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  est la Moyenne,
- $M = \begin{cases} x^{(p+1)} & \text{if } n = 2p + 1 \\ [x^{(p)} + x^{(p+1)}] / 2 & \text{if } n = 2p \end{cases}$  est la Médiane.

Moyenne  $\bar{x}$  et Médiane  $M$ , pondérées par l'inverse de leurs variances  $V(\bar{x})$  et  $V(M)$ , définissent la

$$\text{Médienne : } \bar{M} = \frac{\frac{\bar{x}}{V(\bar{x})} + \frac{M}{V(M)}}{\frac{1}{V(\bar{x})} + \frac{1}{V(M)}} = \frac{V(M)\bar{x} + V(\bar{x})M}{V(\bar{x}) + V(M)},$$

qui peut s'écrire aussi de la façon suivante :  $\bar{M} = (1 - C)\bar{x} + CM$  avec  $C = \frac{V(\bar{x})}{V(\bar{x}) + V(M)}$

Les poids des deux normes peuvent être calculés en simulant les changements dans les distributions. Jackknife et bootstrap pourraient convenir [SHA 1995]. Le bootstrap, mis au point par Efron [EFR1993], a été choisi car il nous assure la normalité de la distribution des résidus (avec un grand nombre de tirages) et permet *in fine* d'estimer une quantité en rééchantillonnant les données. L'algorithme fonctionne comme suit.

- A partir de l'échantillon initial  $X = \{x_1, x_2, \dots, x_N\}$ , nous générons un grand nombre  $B$  de sous-échantillons aléatoires  $\{X^{*1}, X^{*2}, \dots, X^{*B}\}$  par tirage au sort avec remise ;
- Nous calculons les Moyennes et les Médianes de tous les sous-échantillons :  $\{\bar{X}^{*1}, \bar{X}^{*2}, \dots, \bar{X}^{*B}\}$  et  $\{med(X^{*1}), med(X^{*2}), \dots, med(X^{*B})\}$  ;
- La variance  $\hat{S}_{\bar{X}}^2$  des  $B$  Moyennes bootstrappées estime la variance inconnue  $V(\bar{X})$ , la variance  $\hat{S}_{med}^2$  des  $B$  Médianes, la quantité  $V(M)$  :

$$\hat{S}_{\bar{X}}^2 = \frac{1}{B-1} \sum_{b=1}^B [\bar{X}^{*b} - m_{\bar{X}}^*]^2 \quad \text{with} \quad m_{\bar{X}}^* = \frac{1}{B} \sum_{b=1}^B \bar{X}^{*b}$$

$$\hat{S}_{med}^2 = \frac{1}{B-1} \sum_{b=1}^B [med(X^{*b}) - m_{med(X)}^*]^2 \quad \text{with} \quad m_{med(X)}^* = \frac{1}{B} \sum_{b=1}^B med(X^{*b})$$

Dès le début du XIXème siècle, Pierre Simon de Laplace a étudié la distribution jointe de la Moyenne et de la Médiane [LAP 1818, STI 1973]. Il proposa une définition différente de ce que nous appelons aujourd'hui (à peine près de deux siècles plus tard !) la Médiennne. Il y avait introduit un terme correcteur de covariance :

$$\bar{M} = (1 - C)\bar{x} - CM \quad \text{avec} \quad C = \frac{V(\bar{x}) - Cov(\bar{x}, M)}{V(\bar{x}) + V(M) - 2Cov(\bar{x}, M)}$$

Comme pour la première version de la Médiennne, la covariance  $Cov(\bar{x}, M)$  peut être estimée par un bootstrap :  $\hat{cov}_{med, \bar{x}} = \frac{1}{B-1} \sum_{b=1}^B [\bar{X}^{*b} - m_{\bar{X}}^*][med(X^{*b}) - m_{med(X)}^*]$

Suite à ces travaux convergents, les propriétés asymptotiques des Médiennes ont été établies [JOS 2001].

## 5. Evaluation de la Médiennne

Nous évaluons maintenant la robustesse de la Médiennne, quantitativement par des simulations, qualitativement par la cartographie. La comparaison ne porte que sur les Médiennes, la Moyenne et la Médiane, en fonction de l'effectif des distributions, les autres estimateurs n'étant pas l'objet de cet article.

Le tableau 1 donne les résultats des simulations. L'*efficacité relative* de l'estimateur, définie comme le rapport entre sa variance et celle du meilleur testé estimateur (dont de nombreux estimateurs robustes non présentés dans cet article) donne une évaluation de la robustesse de l'estimateur [HOA 1983]. Par exemple, pour une loi de Gauss de 10 individus, l'efficacité de la Moyenne est de 100 % (c'est elle la plus robuste) alors que celle de la Médiane est à 72,3 %. Chaque estimateur est testé sur une série de distributions types à l'aide de 1000 simulations. Pour chacun, on calcule le minimum (MIN) et l'écart-type (ECT) des efficacités relatives de la série. Un estimateur robuste possède un MIN élevé et un faible ECT.

Les résultats montrent que les Médiennes sont globalement toujours plus robustes que la Moyenne, même si localement la Moyenne reste souveraine pour une distribution gaussienne. La Médiennne de Laplace donne systématiquement de meilleurs résultats que la Médiane, la Médiennne sans covariance seulement pour les distributions d'effectifs supérieurs à 100.

L'analyse des cartes de pullulation du campagnol terrestre à l'aide de l'environnement d'analyse statistique exploratoire XlispStat [TIE 1990] révèle la pertinence de la Médienne (Figure 1). Les données brutes ordinales décrivent des niveaux de pullulation du campagnol terrestre dans les communes du Doubs. Les quatre autres fenêtres montrent les résultats d'un filtre spatial par contiguïté d'ordre deux selon les différents estimateurs. On voit que le filtre moyen lisse fortement l'information, le filtre médian délimite des zones distinctes. Les deux filtres médians ont des comportements très similaires. Dans les secteurs géographiques de faible pullulation, ils "rabotent" comme le filtre médian. Dans la partie centrale de la carte, où règne davantage d'hétérogénéité, on obtient des paliers soulignant les gradients. Le bas de la carte, correspondant au massif du Jura, présente une situation intermédiaire, due en partie à la plus grande surface des communes et à un effet de bordure.

Estimateur	Effectif	Gauss	One-out	One-wild	Cont5	Cont10	Dexp	Logistic	Slash	Cauchy	MIN	ECT
Moyenne	10	100	71,9	11,8	85,6	75,7	70,1	94	0	0	0	40,8
Médiane	10	72,3	81,1	76,6	80,1	84	96,7	83,9	90,1	91,8	72,3	7,7
Médienne	10	93,7	87,1	66,5	92,3	90,3	94,6	97,8	90,1	91,9	66,5	9,1
Médienne L	10	100	87,5	77,6	92,4	90,7	98,1	98,4	90,1	91,9	77,6	6,9
Moyenne	20	100	82,5	19,2	84	73,8	63,8	92,8	0	0	0	39,9
Médiane	20	68,1	73,4	70,9	74,5	78	95,7	79,8	85,8	89,5	68,1	9,2
Médienne	20	92,9	88,3	64,4	89,7	87,3	92,6	96,4	85,8	89,4	64,4	9,2
Médienne L	20	100	88,7	71	90,1	87,4	97,1	97,3	85,8	89,4	71	8,6
Moyenne	50	100	92,1	35,6	82,5	72,2	58,5	92	0	0	0	38,9
Médiane	50	65,5	68,1	67	70,5	73,7	96	77,2	82,1	87,3	65,5	10,3
Médienne	50	92,3	90,5	67,1	87,5	84,6	91	95,4	82,2	87,3	67,1	8,3
Médienne L	50	99,9	93,7	69,9	88,2	84,6	96,7	96,5	82,2	87,3	69,9	9,3
Moyenne	100	100	96,1	52,2	82,3	71,4	55,9	92	0	0	0	38,4
Médiane	100	64,6	66,1	65,8	69,4	72,2	96,8	76,4	80,5	86,4	64,6	10,9
Médienne	100	92,2	91,5	73	86,8	83,5	90,5	95,2	80,3	86,4	73	6,9
Médienne L	100	100	96,7	73,6	87,6	83,5	97,3	96,4	80,3	86,4	73,6	9
Moyenne	1000	100	100	91,8	82,2	71,1	51,7	91,8	0	0	0	40,1
Médiane	1000	63,6	64,2	64,3	68,7	71	98,5	75,5	79,3	85,6	63,6	11,7
Médienne	1000	91,9	92,3	89	86,6	82,8	90	95,4	79,3	85,6	79,3	5,1
Médienne L	1000	99,9	100	93,1	87,5	82,8	98,7	96,7	79,3	85,6	79,3	7,9

Tableau 1. Efficacité relative (robustesse) des estimateurs selon l'effectif des distributions et différentes lois

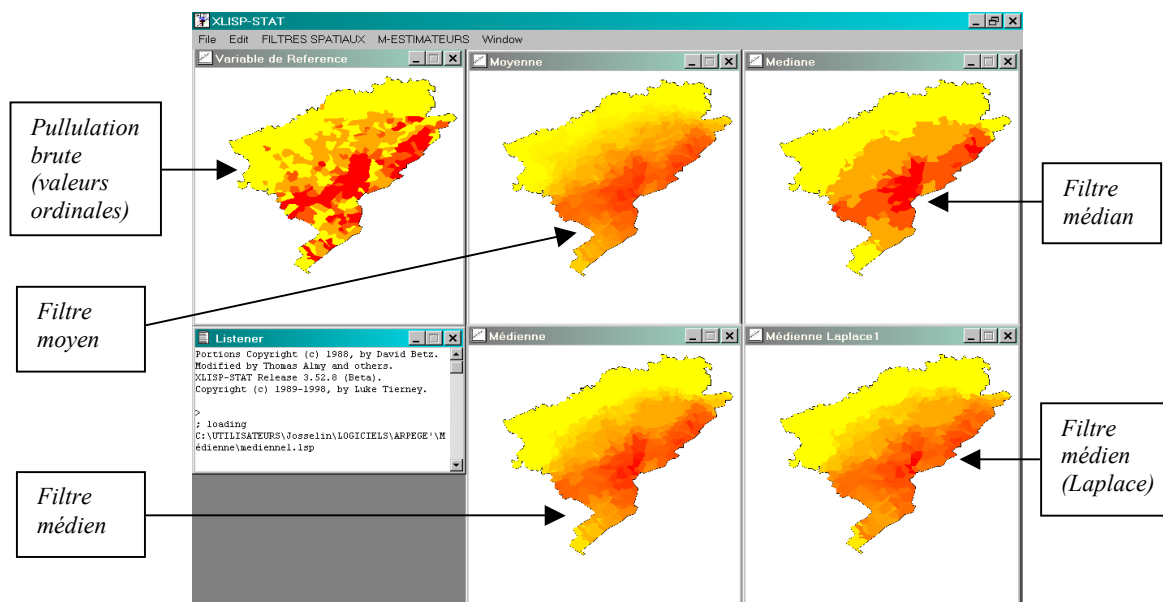


Figure 1. La pullulation du campagnol terrestre dans le Doubs : effet de différents filtres spatiaux d'ordre 2

## 6. Conclusion

Cet article démontre les potentialités importantes de la Médiennne. Ses performances, en termes de robustesse, dépassent celles de la Moyenne et de la Médiane. Le filtre Médien contient, tel qu'il est conçu, une double compétence : lisser ou filtrer. Il est auto-adaptatif, fonction de la distribution statistique qu'il rencontre. Il est conforme à ce que nous recherchions initialement. La version de Laplace améliore sensiblement ses propriétés.

Reste cependant à évaluer plus en profondeur le comportement du filtre médien et à évaluer la Médiennne face aux autres estimateurs robustes de valeurs centrales. Ces travaux sont en cours [JOS 2002].

## 7. Bibliographie

- [AND 1972] ANDREWS D.F., P.J. BICKEL, F.R. HAMPEL, P.J. HUBER, W.H. ROGERS et J.W. TUKEY, *Robust Estimates of Location*, Princeton University Press, Princeton, NJ, 1972.
- [ANS 1995] ANSELIN L., "Local indicators of spatial association, LISA", *Geographical Analysis*, n° 27, pp. 93-115, 1995.
- [CLI 1973] CLIFF AD et ORD JK, *Spatial Autocorrelation*, Pion, London, 1973.
- [COC 1995] COCQUERET JP. et PHILIPP S., *Analyse d'images : filtrage et segmentation*, Masson, Paris, 1995.
- [EFR 1993] EFRON B. et TIBSHIRANI R. J., *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman & Hall, New-York, 1993.
- [FOT 2000] FOTHERINGHAM S., BRUNSDON C., CHARLTON M., 2000 : *Quantitative geography, perspectives on spatial data analysis*, Sage Publications, London, 270 p.
- [GIR 1999] GIRARD M.-C. et GIRARD C, *Traitement des données en télédétection*, Dunod, Paris, 1999.
- [HAM 1986] HAMPEL F., RONCHETTI E., ROUSSEEUW P. et STAHEL W., *Robust Statistics. The approach based on influence functions*, Wiley, New York, 1986.
- [HOA 1983] HOAGLIN D., MOSTELLER F. et TUKEY J.W., *Understanding Robust and Exploratory Data Analysis*, Wiley Series in probability and mathematical statistics, New-York 1983.
- [HUB 1981] HUBER P., *Robust Statistics*, Wiley, New York, 1981.
- [JOS 2000] JOSSELIN D., 2000, "Méthodologies d'analyse exploratoire des données géographiques : tout centrer sur les distributions statistiques et spatiales et les liens dynamiques", *Actes du colloque PRISM-INRETS sur le Data Mining Spatial*, 24-25 février 2000, 14 p.
- [JOS 2001] JOSSELIN D. et LADIRAY D., soumis en Décembre 2001, "Combining L1 and L2 Norms for a more Robust Spatial Analysis : the Meadian Attitude", European Colloquium on Theoretical and Quantitative Geography, Saint-Valery en-Caux September, *Cybergeo*, 2001.
- [JOS 2002] JOSSELIN D. et LADIRAY D., "Back to L-estimator : the Meadian", Working Paper, 15 p, 2002.
- [LAP 1818] LAPLACE P. S. de, *Deuxième supplément à la Théorie Analytique des Probabilités*, Courcier, Paris, 1818.
- [LEC 1987] LECOUTRE J.-P. et TASSI P., *Statistique non paramétrique et robustesse*, Economica, Paris, 1987.
- [SHA 1995] SHAO, J. et TU, D., *The Jackknife and Bootstrap*, XVII, 516 pp, Series: Springer Series in Statistics, Springer-Verlag, 1995.
- [STI 1973] STIGLER, S. M., "Laplace, Fisher, and the discovery of the concept of sufficiency", *Biometrika*, volume 60, n°3, 439-445, 1973.
- [TIE 1990] TIERNEY L., *Lisp-Stat, an object oriented environment for statistical computing and dynamic graphics*, Wiley-Interscience Publication, John Wiley and Sons, NewYork, 1990.
- [WIL 2001] WILCOX, R.R., *Fundamentals of Modern Statistical Methods, Substantially Improving Power and Accuracy*. XIII, 258 p Springer-Verlag, 2001.