

ESTIMATION DE L'INTENSITÉ CIRCULAIRE DES VOYAGEURS INTERNATIONAUX PAR LA MÉTHODE DU NOYAU

Stéphane Tremblay¹

ABSTRACT

K.V. Mardia published a comprehensive account of circular data analysis in 1972. In 1989, N.I. Fisher proposed an estimator of circular density based on kernel method previously developed by M. Rosenblatt in 1956. With circular data, the standard kernel density estimation cannot be used. Applying the wrapping property to the kernel allows one to estimate circular density. In situations where the observations are subject to cyclical variations, the estimator can be used to analyse the underlying cyclical effects. Moreover, the smoothing property of this estimator could remove local fluctuations caused by certain external events such as holidays, strikes, and other special events. We apply this method of estimation in analysing the volume of United States residents entering Canada at St-Stephen landport in 1995. In this case, the local fluctuations could be related to weekends, holidays, special social or cultural events etc., that may effect the volume of travellers. It will be shown that the potential for adaptation, the graphical approach, and the simplicity of interpretation of this estimator make it a useful tool for statistical analysis.

KEYS WORDS: Kernel density estimator; Intensity function; circular density; circular data; smoothing.

RÉSUMÉ

En 1972, K.V. Mardia a publié un compte-rendu détaillé des contributions sur l'analyse des données circulaires. En 1989, N.I. Fisher a proposé un estimateur de densité circulaire basé sur la méthode du noyau développé précédemment par M. Rosenblatt en 1956. Avec des données circulaires, l'estimation de densité standard par le noyau ne peut pas être utilisée. En appliquant la propriété d'enroulement au noyau, nous pouvons estimer la densité circulaire. Dans les situations où les observations sont sujettes à des variations cycliques, l'estimateur peut être utilisé pour analyser les effets cycliques. De plus, la propriété de lissage de cet estimateur peut atténuer les fluctuations locales causées par des événements extérieurs tels que les jours fériés, les grèves et les autres événements spéciaux. Nous appliquons cette méthode d'estimation dans l'analyse du volume des résidents américains qui entrent au Canada par le poste douanier de St-Stephen en 1995. Dans ce cas, les fluctuations locales peuvent être associées aux fins de semaine, aux jours fériés, aux événements sociaux ou culturels, etc., qui peuvent affecter le nombre de voyageurs. On montrera que le potentiel d'adaptation, l'approche graphique et la simplicité de l'interprétation de cet estimateur, font de lui un outil utile pour l'analyse statistique.

MOTS CLÉS: Estimateur de densité à noyau; intensité circulaire; densité circulaire; observations circulaires; lissage.

1. INTRODUCTION

Il est maintenant bien connu que l'estimateur à noyau est un estimateur non paramétrique de densité robuste dont les champs d'application ne cessent de se diversifier. Sa flexibilité et sa simplicité suscitent un intérêt croissant en recherche appliquée. Plusieurs exemples illustrant les forces de cet estimateur sont montrés dans les monographies suivantes : Silverman (1986), Härdle (1990), Scott (1992) et Wand & Jones (1995).

Soit X_1, \dots, X_n , un échantillon de n observations indépendantes et identiquement distribuées provenant d'une loi absolument continue de densité f sur \mathbb{R} . L'estimateur à noyau usuel de densité est alors défini par

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad (1)$$

où K est le noyau étant généralement une densité symétrique, $K_h(z) = h^{-1}K(h^{-1}z)$ et $h > 0$ est le paramètre de lissage. Par contre, lorsque les observations sont dites

¹ Stéphane Tremblay, Division des méthodes d'enquêtes des ménages, Statistique Canada, 16-K édifice R.H. Coats, Ottawa (Ontario), Canada, K1A 0T6, E-Mail: tremste@statcan.ca.

circulaires, *i.e.* possédant un phénomène périodique, l'estimateur (1) doit subir quelques modifications (voir section 2). Nous aurons alors un estimateur à noyau de densité circulaire. À la section 3, une brève description de l'ensemble de données sera donnée et aux sections 4 et 5, l'estimateur circulaire multiplié par le nombre d'observations sera utilisé pour illustrer le volume américains. Finalement à la section 7, nous brosserons un survol de la littérature faisant une revue des applications possibles de la méthode du noyau et nous concluons cette article.

2. ESTIMATEUR CIRCULAIRE

On définit une observation circulaire comme étant une valeur d'un vecteur aléatoire Z , prenant ses valeurs sur la circonférence du cercle unité $x^2 + y^2 = 1$ dans le plan. Nous identifions les valeurs possibles de Z avec des angles mesurés dans le sens antihoraire à partir de l'abscisse positive. De cette représentation, si Z possède une densité f_c , dite circulaire, sur $[0, 2\pi]$, nous pouvons écrire la fonction de répartition F_c de Z sous la forme

$$F_c(\theta) = P(0 \leq Z \leq \theta) = \int_0^\theta f_c(z) dz, \quad 0 \leq \theta \leq 2\pi. \quad (2)$$

Étant donné que les données circulaires apparaissent dans les situations où un phénomène périodique, *i.e.* cyclique, est présent, il est naturel de poser les conditions que $f_c(0) = f_c(2\pi)$, et $f_c'(0_+), f_c'(2\pi_-)$ existent et sont égales. On dira qu'une densité sur $[0, 2\pi]$ ayant ces propriétés est une *densité circulaire*. Nous considérons $\theta_1, \dots, \theta_n$, un échantillon de n angles d'une loi absolument continue de densité circulaire f_c sur $[0, 2\pi]$. Dans cette situation, l'estimateur à noyau (1) est incapable de respecter les conditions susmentionnées d'une densité circulaire. Il devient alors nécessaire d'apporter certaines modifications à cet estimateur.

À toute loi définie sur \mathbb{R} , on peut associer sa loi dite *enroulée* (voir Mardia (1972, pp. 53-54)). De manière précise, si x est une variable aléatoire dans \mathbb{R} , on peut lui faire correspondre la variable

$$x^e = x \pmod{2\pi}.$$

Si f est la fonction de densité de x et que sa dérivée existe et est continue, son enroulement est alors défini par la densité de x^e :

$$f^e(\theta) = \sum_{l=-\infty}^{\infty} f(\theta + 2\pi l), \quad 0 \leq \theta \leq 2\pi. \quad (3)$$

En d'autres termes, la propriété d'enroulement nous dit

que l'enroulement de densités sur la droite des réels produit des densités circulaires. En appliquant la propriété d'enroulement (3) au noyau K , il est alors naturel d'estimer $f^e(\theta)$ par

$$\hat{f}^e(x) = \frac{1}{n} \sum_{i=1}^n K_h^e(\theta - \theta_i), \quad 0 \leq \theta \leq 2\pi, \quad (4)$$

où K^e est le noyau K enroulé défini par

$$K^e(\theta) = \sum_{l=-\infty}^{\infty} K(\theta + 2\pi l), \quad 0 \leq \theta \leq 2\pi. \quad (5)$$

Étant donné que cet estimateur contient une sommation infinie, nous utilisons en pratique l'estimateur suivant:

$$\tilde{f}^e(\theta) = \sum_{l=-l_h}^{l_h} \frac{1}{n} \sum_{i=1}^n K_h(\theta + 2\pi l - \theta_i) \quad (6)$$

pour un l_h approprié. Nous distinguons deux cas selon que le support de K soit borné ou non borné. Étant donné qu'il est très rare en pratique de rencontrer $h \geq 2\pi$, nous limitons le choix du paramètre de lissage à $0 < h < 2\pi$. Ainsi d'après Tremblay (1997, pp. 57-59), il est possible de montrer qu'avec $l_h \geq 1$, nous avons $\hat{f}^e = \tilde{f}^e$ lorsque le support de K est borné. Lorsque le support de K est non borné, nous ne pouvons avoir une égalité entre ces deux estimateurs. En revanche, en utilisant $l_h = 4$ nous avons une approximation raisonnablement bonne.

3. CHOIX DE K ET h

Anderson (1969), par des simulations, et Scott (1992), en étudiant l'efficacité relative asymptotique des noyaux usuels, ont montré que le choix du noyau K est plus ou moins important et qu'il pouvait être basé sur des arguments tels sa différentiabilité, sa facilité de calculs, *etc.* Par contre, le choix du paramètre h est beaucoup plus critique (voir Tremblay (1997)).

En utilisant le noyau quartique $K(x) = 0.9375(1 - x^2)^2$, pour $|x| \leq 1$ et zéro autrement,

$$h = \sqrt{7} \hat{\sigma} n^{-1/5} \quad (7)$$

Fisher (1989) suggère le paramètre de lissage suivant: où n est le nombre d'observation choisi en fonction de

la modalité de la densité², $\hat{\sigma}$ est l'estimateur du maximum de vraisemblance du paramètre de concentration d'une distribution de van Mises (voir Mardia (1972)) et $\sqrt{7}$ est justifié par l'utilisation du noyau quartique.

4. ENSEMBLE DE DONNÉES

L'ensemble de données utilisé dans ce document est le volume journalier d'Américains qui sont restés au Canada une nuit ou plus et qui sont entrés par le poste douanier de St-Stephen (NB) en 1995. Plusieurs raisons ont motivé ce choix. Premièrement, avec un volume 134957 Américains passant une nuit ou plus au Canada, St-Stephens est le plus important poste frontalier des Provinces maritimes. Deuxièmement, ce poste frontalier est situé à quelques kilomètres de Fredericton, la capitale provinciale du Nouveau-Brunswick où avait lieu la conférence. Finalement, ce poste frontalier semble avoir une distribution du volume trimestriel semblable à plusieurs autres à travers le Canada, *i.e.* avec un volume estival relativement haut comparativement à celui de l'hiver. Plusieurs applications dont l'étude de la distribution de questionnaires sont possibles.

Ces données sont collectées par Revenu Canada et sont transmises à Statistique Canada (SC). Ensuite, le personnel de Statistique Canada responsable de l'Enquête sur les voyages internationaux (EVI) s'occupe de gérer, de compiler et de publier ces données.

5. DISCUSSION

Par le biais de l'EVI, SC détermine les périodes d'échantillonnage et le nombre de questionnaires à distribuer selon le volume trimestriel de l'année précédente. En outre, SC désire conserver une fraction de sondage constante au cours des trimestres d'une même année.

Une méthode de distribution de questionnaires de ce type est considérée "bonne" lorsque la fraction sondage est relativement constante. Dans cette exemple, le volume estimé, appelé *l'intensité estimée du volume*, doit être assez souple pour illustrer les tendances importantes et assez rigide pour lisser les fluctuations locales. Dans cette section, nous comparerons graphiquement la méthode actuelle de distribution de questionnaires à une nouvelle méthode utilisant l'estimateur à noyau.

La méthode de distribution de questionnaires

actuelle est proportionnelle au volume journalier moyen du trimestre de l'année précédente. Entre autres, cette méthode suppose que les volumes journaliers dans ce trimestre sont relativement uniformes. Graphiquement, nous pouvons illustrer ce type d'intensité estimée du volume à l'aide d'un histogramme avec quatre tiges (une par trimestre) dont la hauteur est le volume journalier moyen du trimestre (voir figure 1). La méthode de distribution de questionnaires proposée est proportionnelle à l'estimateur à noyau de l'intensité suggéré par Diggle (1985). Ce nouvel estimateur est le résultat du produit entre le volume annuel, $n = 134957$, et de l'estimateur (6) (voir figure 2). Remarquons que l'illustration circulaire met davantage en lumière la tendance des volumes journaliers que l'illustration linéaire (figures 1 et 2).

6. CONCLUSION

Outre l'application aux observations circulaires, l'estimateur à noyau peut être utilisé dans plusieurs domaines de la statistique dont l'analyse des données et l'inférence statistique. Entre autres, les auteurs Silverman (1986), Härdle (1991), Scott (1992), Wand & Jones (1995) et Tremblay (1997) présentent une revue détaillée et relativement complète des applications couramment utilisées.

Dans cet article, nous avons présenté l'estimateur à noyau comme un estimateur non paramétrique simple, intuitif et robuste vis-à-vis différentes situations. La capacité d'analyse et la simplicité d'interprétation qu'offre cette méthode font de celle-ci un outil fort attrayant pour tous ceux qui s'intéressent aux applications statistiques.

D'après ces deux figures, il semble évident que la méthode proposée respecte mieux la fraction de sondage que la méthode actuelle. Notons que l'illustration circulaire des données et des intensités estimées est possible grâce au cycle annuel de ce type de volume. Étant donné que le volume journalier est relativement stable au cours des trimestres 1 et 4, la méthode actuelle semble raisonnable (voir figure 1). En revanche, pendant les trimestres les plus occupés de l'année, *i.e.* le deuxième et le troisième, la méthode actuelle surestime la fraction de sondage au début du 2-ième trimestre, la sous-estime entre le 2-ième et le 3-ième et finalement, la surestime à nouveau à la fin du 3-ième trimestre.

² Si f_c est unimodal, n est le nombre d'observations dans l'échantillon sinon n est le nombre d'observations dans le plus important mode de la densité.

Quant à la méthode proposée, elle semble mieux suivre la tendance des volumes journaliers et par conséquent, mieux respecter la fraction de sondage. Nous remarquons que même si la variabilité des volumes journaliers au cours de la période estivale est grande, l'intensité estimée par la méthode du noyau demeure relativement stable. Par conséquent, il peut résister aux fluctuations dues à des événements extraordinaires tels des jours fériés, des grèves, des catastrophes météorologiques, etc.

RÉFÉRENCES

Anderson, G.D. (1969). *A Comparison of Methods for Estimating a Probability Density Function*, Ph.D. thesis, University of Washington.

Diggle, P.J. (1985). A kernel method for smoothing point process data, *Applied Statistics* (34): 138-147.

Fisher, N.I. (1989). Smoothing a sample of circular data, *Journal of Structural Geology* (11): 775-778

Fisher, N.I. (1993). *Statistical Analysis of Circular Data*, Cambridge Press.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S*, New-York: Springer-Verlag.

Mardia, K.V. (1972). *Statistics of Directional Data*, London: Academic Press.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New-York.

Silverman, B.W. (1989). *Density Estimation for Statistics and Data Analysis*, New-York: Chapman and Hall.

Tremblay, S. (1997). *La méthode du noyau et quelques-unes de ses applications*, Mémoire de maîtrise, Université Laval.

Wand, M.P. et Jones, M.C. (1995). *Kernel Smoothing*, London UK: Chapman and Hall.

Figure 1: Illustration de la méthode actuelle: intensité estimée par l'histogramme du volume d'Américains au poste St-Stephens (N.-B.) en 1996.

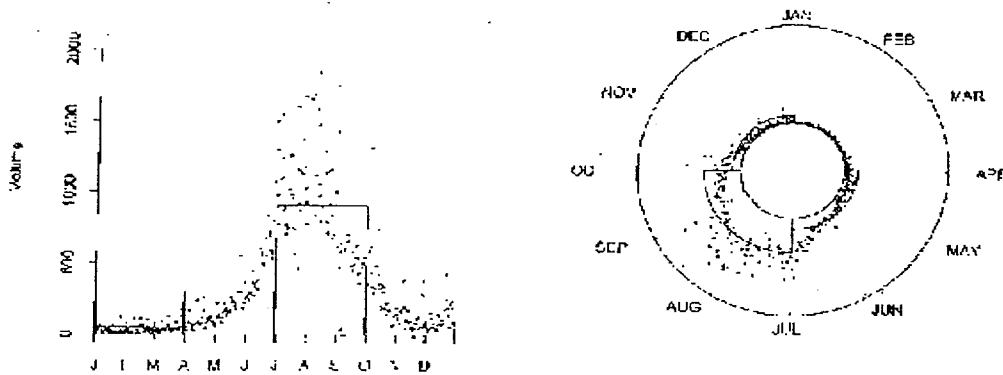


Figure 2: Illustration de la méthode proposée: intensité estimée par l'estimateur à noyau du volume d'Américains au poste St-Stephens (N.-B.) en 1996.

