Journal of Mathematics and Statistics

Original research paper

A Suggested Nonparametric Bivariate Logistic Density Estimator with Application on the Productivity of Egyptian Wheat during 2019/2020

Samah M. Abo-El-Hadid

Department of Mathematics, Insurance and Applied Statistics,
Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

Article history
Received: 14-04-2021
Revised: 17-05-2021
Accepted: 21-05-2021

Email: s_aboelhadid@yahoo.com
Samah_2999@yahoo.com

Abstract: In this study, the nonparametric standard logistic density estimator, introduced by Abo-El-Hadid (2018), is extended to the bivariate case. The multiplicative standard logistic distribution is used as a kernel function to derive the bivariate kernel estimator. The statistical properties of the resulting estimator are studied, which are: The asymptotic bias, variance, Mean Squared Error (*MSE*) and Integrated Mean Squared Error (*IMSE*); also, the optimal bandwidth is obtained. A simulation study is introduced to investigate the performance of the proposed estimator with other estimators. We also apply the proposed estimator to a real data set to estimate the bivariate density of the planted and productive areas of wheat in Egypt.

Keywords: Joint Kernel Density Estimator, Bivariate Logistic Kernel, Optimal Bandwidth

Introduction

In many situations, there are, more than one random variable of interest; hence we are needed to extend the density function of one random variable to those of two or more random variables.

Rosenblatt (1956) introduced a univariate nonparametric estimator of the density function $f(x)$, which called the kernel density estimator:

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \quad (1)$$

where, n is the sample size; $K(\blacksquare)$ and h are the univariate kernel function and the bandwidth respectively, where the univariate kernel function $K(\blacksquare)$ is assumed to be a density function. This kernel density estimator in (1) is extended to the multivariate case by Epanechnikov (1969).

Duong and Hazelton (2003) used plug-in methods for selecting the bandwidth matrix for bivariate kernel density estimation. The performance of their methodology is compared with existing plugin techniques via numerical study.

Santhosh and Srinivas (2013) used Diffusion process based on adaptive kernel to model joint distributions of peak flow and volume that characterize a flood data extracted from daily streamflow records pertaining to stations in India, United Kingdom, Canada and United States. The performance of the D-kernel is compared with that of kernel density estimation using Gaussian kernel, The D-kernel is shown to be effective when compared to Gaussian kernel.

Cañón-Tapia (2013) suggested using the bivariate Gauss kernel to study the spatial distribution of volcanic vents. The suggested bivariate Gauss kernel is compared with Fisher kernel and it is found that both kernels can be used to obtain the same general description of volcanic distribution.

Bandyopadhyay and Modak (2018) introduce estimators based on the product of a univariate classical kernel and a univariate gamma kernel and compare their performances in terms of the mean integrated squared error. Two astronomical data sets are used to demonstrate the applicability of this estimator.

The rest of this paper is organised as follows: In section 2 the suggested bivariate logistic kernel estimator and its statistical properties are introduced. In section 3 the optimal bandwidth is obtained. a simulation study is introduced in section 4. Real data application is introduced in section 5.

The Suggested Bivariate Logistic Density Estimator

Let the joint probability density function of the two random variables X_1, X_2 be $f(x_1, x_2)$. The bivariate kernel density estimator given by Epanechnikov (1969) takes the form:

$$\hat{f}\left(x_1,x_2\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\mathcal{K}\left(\frac{x_1-X_{i1}}{h},\frac{x_2-X_{i2}}{h}\right) \tag{2}$$

where, h is the smoothing parameter under the assumption that the bandwidth is the same for both X_1, X_2. $\mathcal{K}$ is bivariate kernel function, which for simplicity considered as a multiplicative kernel (Silverman, 1986):

$$\mathcal{K}(u)=K\left(u_1\right)\cdot K\left(u_2\right) \tag{3}$$

where, K denotes a univariate kernel function. Then:

$$\hat{f}\left(x_1,x_2\right)=\frac{1}{nh^2}\sum_{i=1}^{n}\left[\prod_{j=1}^{2}K\left(\frac{x_j-X_{ij}}{h}\right)\right] \tag{4}$$

$$\therefore\ \hat{f}\left(x_1,x_2\right)=\frac{1}{nh^2}\sum_{i=1}^{n}\left[K\left(\frac{x_1-X_{i1}}{h}\right)\cdot K\left(\frac{x_2-X_{i2}}{h}\right)\right] \tag{5}$$

Using the multidimensional form of Taylor's theorem for the υ th order of kernel function, Epanechnikov (1969) proved that the bias of $\hat{f}\left(x_1,x_2\right)$ is as follows (Härdle *et al*., 2004):

$$Bias\left[\hat{f}\left(x_1,x_2\right)\right]=\frac{h^{\upsilon}\sigma_{\mathcal{K}}^{\upsilon}}{\upsilon!}\sum_{j=1}^{q}\frac{\partial^{\upsilon}}{\partial x_j^2}f\left(x_1,\cdots,x_q\right)+o\left(h^2\right) \tag{6}$$

Then for the 2nd order kernel function:

$$\begin{aligned}&Bias\left[\hat{f}\left(x_1,x_2\right)\right]\\&=\frac{h^2\sigma_{\mathcal{K}}^2}{2}\left[\frac{\partial^2}{\partial x_1^2}f\left(x_1,x_2\right)+\frac{\partial^2}{\partial x_2^2}f\left(x_1,x_2\right)\right]+o\left(h^2\right)\end{aligned} \tag{7}$$

where, $o(h^2)$ is a higher order term than 2 of h.

Also, Epanechnikov (1969) found that the variance takes the following form:

$$Var\left[\hat{f}\left(x_1,x_2\right)\right]=\frac{f\left(x_1,x_2\right)R(\mathcal{K})}{nh^2}+o\left(\frac{1}{n}\right) \tag{8}$$

$$\therefore Var\left[\hat{f}\left(x_1,x_2\right)\right]=\frac{f\left(x_1,x_2\right)\left[R(\mathcal{K})\right]^2}{nh^2}+o\left(\frac{1}{n}\right) \tag{9}$$

where, $R(K)=\int K^2(\blacksquare)d\blacksquare$.

In this study, we suggest use the standard logistic distribution as a kernel function. The univariate standard logistic distribution takes the form (Evans *et al*., 1993):

$$K_0\left(v\right)=\frac{e^{-v}}{\left(1+e^{-v}\right)^2},\quad -\infty\le v\le\infty \tag{10}$$

Where:

$$E\left(v\right)=0 \tag{11}$$

$$\sigma_K^2=\mathrm{var}\left(v\right)=\frac{\pi^2}{3} \tag{12}$$

Also as proved by Abo-El-Hadid (2018):

$$R\left(K\right)=\int_{-\infty}^{\infty}K^2\left(v\right)dv=\frac{1}{6} \tag{13}$$

Then the suggested multiplicative kernel $\mathcal{K}$ is as follows:

$$\mathcal{K}(u)=K\left(u_1\right)\cdot K\left(u_2\right)=\frac{e^{-u_1}\cdot e^{-u_2}}{\left(1+e^{-u_1}\right)^2\cdot\left(1+e^{-u_2}\right)^2} \tag{14}$$

Then $E(u) = 0$:

$$\sigma_{\mathcal{K}}^2=\mathrm{var}\left(u\right)=\frac{\pi^4}{9} \tag{15}$$

$$R\left(\mathcal{K}\right)=R^2\left(K\right)=\frac{1}{36} \tag{16}$$

Using the multiplicative kernel in (14), the suggested bivariate logistic kernel density estimator is as follows:

$$\hat{f}\left(x_1,x_2\right)=\frac{1}{nh^2}\sum\nolimits_{i=1}^{n}\left[\frac{e^{-\left(\frac{x_1-X_{i1}}{h}\right)}}{\left[1+e^{-\left(\frac{x_1-X_{i1}}{h}\right)}\right]^2}\cdot\frac{e^{-\left(\frac{x_2-X_{i2}}{h}\right)}}{\left[1+e^{-\left(\frac{x_2-X_{i2}}{h}\right)}\right]^2}\right] \tag{17}$$

The bias of the above estimator is obtained by substituting from Equation (15) into Equation (7):

$$Bias\left[\hat{f}\left(x_1,x_2\right)\right]\approx\frac{\pi^4h^2}{18}\left[\frac{\partial^2}{\partial x_1^2}f\left(x_1,x_2\right)+\frac{\partial^2}{\partial x_2^2}f\left(x_1,x_2\right)\right] \tag{18}$$

And substitute from Equation (16) into Equation (8) yields that the approximated variance is:

$$Var\left[\hat{f}\left(x_1,x_2\right)\right]\approx\frac{f\left(x_1,x_2\right)}{36nh^2} \tag{19}$$

Combining Equation (18) and Equation (19), then the asymptotic mean squared error MSE is:

$$MSE\left[\hat{f}\left(x_1,x_2\right)\right]=\frac{f\left(x_1,x_2\right)}{36nh^2}+\frac{\pi^8h^4}{324}\left[\frac{\partial^2}{\partial x_1^2}f\left(x_1,x_2\right)+\frac{\partial^2}{\partial x_2^2}f\left(x_1,x_2\right)\right]^2$$

Then the integrated mean squared error IMSE is as follows:

$$IMSE\left[\hat{f}\left(x_1,x_2\right)\right]=\frac{1}{36nh^2}+\int\int_{-\infty}^{\infty}\frac{\pi^8h^4}{324}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2 \quad (20)$$

The Optimal Bandwidth

In this section, the optimal smoothing parameter is derived, that minimize the *IMSE*:

$$\frac{\partial IMSE\left[\hat{f}\left(x\right)\right]}{\partial h}=\frac{-2}{36nh^3}+\frac{4h^3\pi^8}{324}\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2=0 \quad (21)$$

Then:

$$\frac{2}{36n}=\frac{4h^6\pi^8}{324}\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2$$

$$h^6=\left[\frac{2}{9}n\pi^8\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2\right]^{-1} \quad (22)$$

$$\therefore h=\left[\frac{2}{9}n\pi^8\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2\right]^{-\frac{1}{6}}$$

The optimal smoothing parameter in (22) depends on the unknown term $\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2$, to overcome this problem, the multiplicative standard logistic distribution is used as a reference distribution. Then:

$$f\left(x\right)=\frac{e^{-x_1}\cdot e^{-x_2}}{\left(1+e^{-x_1}\right)^2\cdot\left(1+e^{-x_2}\right)^2},-\infty\le x_1,x_2\le\infty \quad (23)$$

Then:

$$\frac{\partial f\left(x_1,x_2\right)}{\partial x_1}=\frac{2e^{-2x_1-x_2}}{\left(1+e^{-x_1}\right)^3\left(1+e^{-x_2}\right)^2}-\frac{e^{-x_1-x_2}}{\left(1+e^{-x_1}\right)^2\left(1+e^{-x_2}\right)^2}$$

$$\therefore\frac{\partial f\left(x_1,x_2\right)}{\partial x_1}=\frac{e^{-x_1-x_2}\left(e^{-x_1}-1\right)}{\left(1+e^{x_1}\right)^3\left(1+e^{x_2}\right)^2} \quad (24)$$

$$\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}=\frac{3e^{2x_1+x_2}\left(e^{x_1}-1\right)}{\left(1+e^{x_1}\right)^4\left(1+e^{x_2}\right)^2}-\frac{e^{2x_1+x_2}}{\left(1+e^{x_1}\right)^3\left(1+e^{x_2}\right)^2}$$

$$-\frac{e^{x_1+x_2}\left(e^{x_1}-1\right)}{\left(1+e^{x_1}\right)^3\left(1+e^{x_2}\right)^2}\therefore\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}=\frac{e^{x_1+x_2}\left(e^{2x_1}-4e^{x_1}+1\right)}{\left(1+e^{x_1}\right)^4\left(1+e^{x_2}\right)^2} \quad (25)$$

$$\frac{\partial f\left(x_1,x_2\right)}{\partial x_1}=\frac{2e^{-x_1-2x_2}}{\left(1+e^{-x_1}\right)^2\left(1+e^{-x_2}\right)^3}-\frac{e^{-x_1-x_2}}{\left(1+e^{-x_1}\right)^2\left(1+e^{-x_2}\right)^2}$$

$$\therefore\frac{\partial f\left(x_1,x_2\right)}{\partial x_1}=\frac{e^{x_1+x_2}\left(e^{x_2}-1\right)}{\left(1+e^{x_1}\right)^2\left(1+e^{x_2}\right)^3} \quad (26)$$

$$\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}=\frac{3e^{x_1+2x_2}\left(e^{x_2}-1\right)}{\left(1+e^{x_1}\right)^2\left(1+e^{x_2}\right)^4}-\frac{e^{x_1+2x_2}}{\left(1+e^{x_1}\right)^2\left(1+e^{x_2}\right)^3}$$

$$-\frac{e^{x_1+x_2}\left(e^{x_2}-1\right)}{\left(1+e^{x_1}\right)^2\left(1+e^{x_2}\right)^3}\therefore\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}=\frac{e^{x_1+x_2}\left(e^{2x_2}-4e^{x_2}+1\right)}{\left(1+e^{x_1}\right)^2\left(1+e^{x_2}\right)^4} \quad (27)$$

then:

$$\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2$$

$$=\int_{-\infty}^{\infty}\frac{-195e^{5x_1}+835e^{4x_1}-1955e^{3x_1}-477e^{2x_1}-224e^{x_1}+32}{3150\left(1+e^{x_1}\right)^7}dx_1 \quad (28)$$

$$\therefore\int\int_{-\infty}^{\infty}\left[\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_1^2}+\frac{\partial^2 f\left(x_1,x_2\right)}{\partial x_2^2}\right]^2 dx_1dx_2=\frac{32}{3150}$$

Substituting (28) into (22), the optimal bandwidth is:

$$h_{opt}=\left[\frac{32}{14195}n\pi^8\right]^{-\frac{1}{6}} \quad (29)$$

Simulation

In this section, the performance of the proposed bivariate logistic kernel estimator is evaluated via simulation. The suggested estimator was compared with:

- The multiplicative Gaussian kernel estimator:

$$K\left(u\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 e^{\left(-\frac{1}{2}u_1^2\right)}\cdot e^{\left(-\frac{1}{2}u_2^2\right)}$$

- The multiplicative Epanechnikov kernel:

$$K\left(u\right)=\left(\frac{3}{4\sqrt{5}}\right)^2\left(1-\frac{u_1^2}{5}\right)\cdot\left(1-\frac{u_2^2}{5}\right)$$

Random samples were generated from: Bivariate logistic distribution (Gumbel, 1961). The size of the random samples is $n\in\{10,50,100,500,1000\}$. Then the following measures of error are computed:

$$Mean\ Squared\ Error\left(MSE\right)=\frac{\sum_{i=1}^{n}\left(f\left(x_{i1},x_{i2}\right)-\hat{f}\left(x_{i1},x_{i2}\right)\right)^2}{n} \quad (30)$$

$$Mean\ Qbsolute\ Error(MAE) = \frac{\sum_{i=1}^{n}\left|f(x_{i1},x_{i2}) - \hat{f}(x_{i1},x_{i2})\right|}{n} \quad (31)$$

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \sum_{i=1}^{n}\left|\frac{f(x_{i1},x_{i2}) - \hat{f}(x_{i1},x_{i2})}{n\cdot f(x_{i1},x_{i2})}\right| \quad (32)$$

The values of the above goodness of fit measures are given in the following Table 1.

From the above error measures in Table 1, it is obvious that the performance of all estimates improved as the sample size increases. It is also noted that the worst estimate according to the criteria of error is the multiplicative Epanechnikov estimator, followed by the multiplicative Gaussian estimator, while the suggested bivariate logistic estimator overcomes the other estimators.

Application

Agriculture is an important sector of the Egyptian economic development sectors. Wheat is a major crop and one of the main pillars of the Egyptian economics. Egypt is the largest consumer and the largest importer of wheat in the world. This wheat is mainly used in the production of Egyptian bread. hence, wheat is a product of utmost importance to Egypt and reforming the bread program is a top priority for the Egyptian government.

Therefore, this study focuses on studying the joint probability density function of wheat cultivated areas in the governorates of Egypt and the productivity of those governorates (27 governorates).

Table 1: Simulation results

n	Estimator	*MSE*	*MAE*	*MAPE*
10	Multiplicative Gaussian	0.159167	0.322847	0.567504
	Multiplicative Epanechnikov	0.175300	0.34694	0.624018
	Suggested logistic	0.156474	0.295652	0.523707
50	Multiplicative Gaussian	0.155581	0.309598	0.563726
	Multiplicative Epanechnikov	0.171997	0.331276	0.623491
	Suggested logistic	0.145774	0.281742	0.512723
100	Multiplicative Gaussian	0.151168	0.308198	0.551917
	Multiplicative Epanechnikov	0.168923	0.329561	0.607196
	Suggested logistic	0.137271	0.271262	0.49482
500	Multiplicative Gaussian	0.140966	0.282875	0.524303
	Multiplicative Epanechnikov	0.155848	0.304107	0.581337
	Suggested logistic	0.125577	0.265113	0.461971
1000	Multiplicative Gaussian	0.00296022	0.0478740	0.474105
	Multiplicative Epanechnikov	0.00554374	0.0521915	0.572888
	Suggested logistic	0.00228290	0.0438053	0.44933

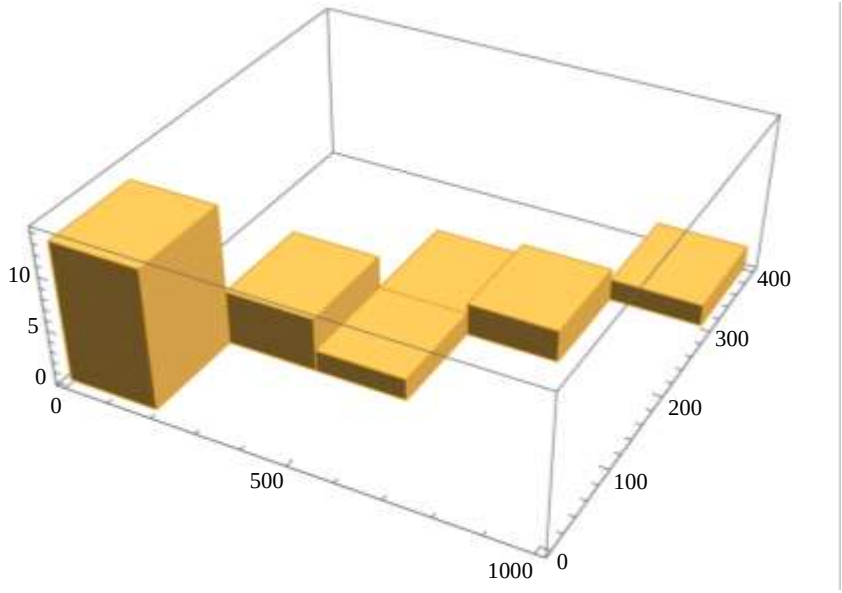

Fig. 1: The histogram of production and planted area of Egyptian wheat by governorate during 2019/2020

The data on the planted area in feddans (feddan = 4200 square meters) and wheat production (in tons per feddans) by governorate are obtained from the central agency for public mobilization and statistics: Egypt in figures agriculture (March 2021). The following graph illustrate the bivariate histogram of planted area and production of wheat together. A bivariate histogram bins the data within rectangles and then shows the count of observations within each rectangle (Fig. 1).

The suggested bivariate logistic kernel estimator is used to estimate the bivariate density function of wheat production and planted area and Fig. 2 introduce the 3D graph of the estimated density

A contour plot is a way of displaying the above 3D plot on a 2D plot. the contour plot shows only two dimensions (the x-axis and the y-axis), the third dimension is defined by the colour. The following figure illustrate the contour plot of the estimated density function and the cumulative function.

From Fig. 3, it is obvious that we have two positively correlated variables, because there is overall tendency of the contour lines to point up and to the right (or down and to the left).

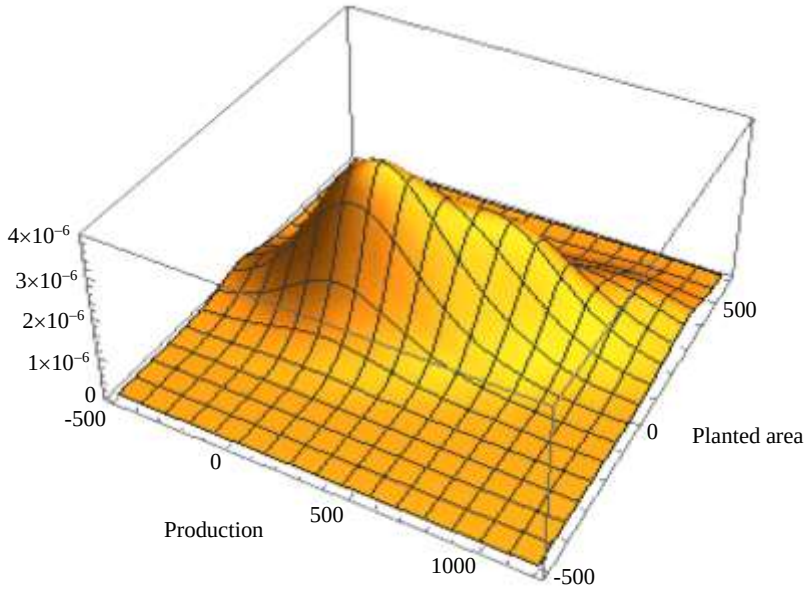

Fig. 2: 3D graph of the estimated bivariate density of production and planted area of Egyptian wheat

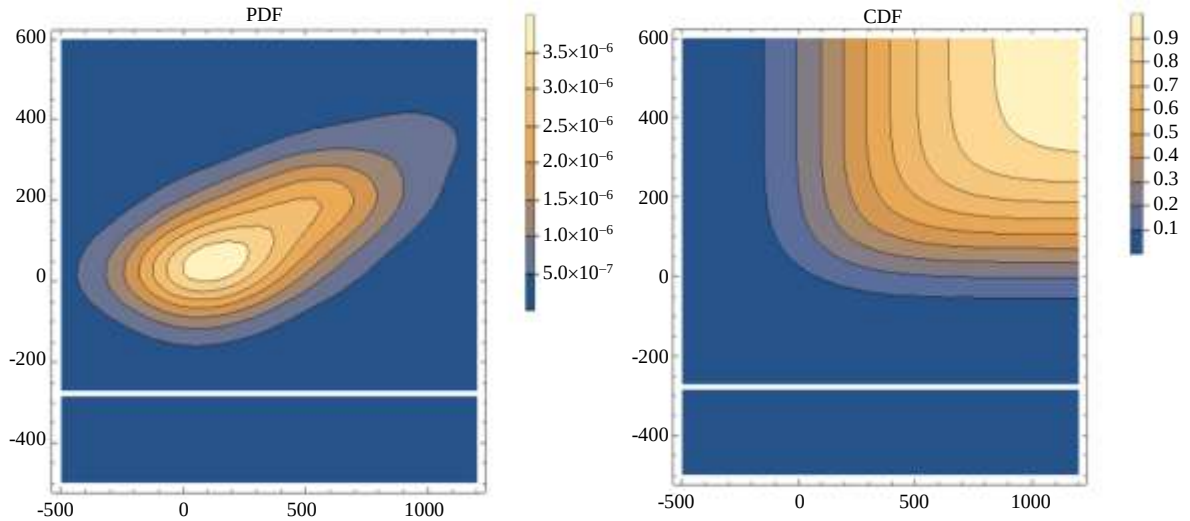

Fig. 3: The PDF and CDF of production and planted area of Egyptian wheat

Conclusion

In this study, the univariate logistic kernel estimator introduced by Abo-El-Hadid (2018) is extended to the bivariate case. The theoretical properties of this proposed estimator are studied. Since the wheat crop is of prime importance to Egypt, due to its use in the manufacture of a basic commodity, which is the "baladi" bread, the bivariate probability density function of wheat yield and the areas planted with it was estimated. We found from the estimated density that there is as positive relationship between the wheat production and the planted areas. Finally, from a simulation study, the new bivariate logistic kernel estimator always outperforms the other estimators.

Ethics

The author confirms that this article is original and contains unpublished material. And the author has read and approved the manuscript and no ethical issues involved.

References

Abo-El-Hadid, S. M. (2018). Logistic Kernel Estimator and Bandwidth Selection for Density Function. International Journal of Contemporary Mathematical Sciences, 13(6), 279-286. https://doi.org/10.12988/ijcms.2018.81133

Bandyopadhyay, U., & Modak, S. (2018). Bivariate density estimation using normal-gamma kernel with application to astronomy. Journal of Applied Probability and Statistics, 13, 23-39. https://arxiv.org/abs/1801.08300

Cañón-Tapia, E. (2013). Volcano clustering determination: Bivariate Gauss vs. Fisher kernels. Journal of Volcanology and Geothermal Research, 258, 203-214. https://doi.org/10.1016/j.jvolgeores.2013.04.015

Duong, T., & Hazelton, M. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics, 15(1), 17-30. https://doi.org/10.1080/10485250306039

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153-158. https://doi.org/10.1137/1114019

Evans, M., Hastings, N., & Peacock, B. (1993). Statistical Distributions. John Wiley and Sons Inc. New York. https://doi.org/10.1002/asm.3150100411

Gumbel, E. J. (1961). Bivariate logistic distributions. Journal of the American Statistical Association, 56(294), 335-349. https://doi.org/10.1080/01621459.1961.10482117

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). Nonparametric and Semi-parametric Models: An Introduction. Springer-Verlag, New York, ISBN-10: 3540207228.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of Density Function. The Annals of Mathematical Statistics, 27, 832-837. https://doi.org/10.1214/aoms/1177728190

Santhosh, D., & Srinivas, V. V. (2013). Bivariate frequency analysis of floods using a diffusion based kernel density estimator. Water Resources Research, 49(12), 8328-8343. https://doi.org/10.1002/2011WR010777

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York, ISBN-10: 0412246201.