Vol. 4 Issue 1 Meteorology Hydrology and Water Management Research and Operational Applications June 2016

The use of kernel estimators to determine the distribution of groundwater level

Andrzej Michalski

Wrocław University of Environmental and Life Sciences, Department of Mathematics, Grunwaldzka 53, 50-357 Wrocław, Poland, e-mail: apm.mich@gmail.com

Abstract. In this paper the problem of non-parametric estimation of the probability density function for hydrological data is considered. For a given random sample $X_1, X_2, ..., X_n$ we define an estimator $\hat{f}_n$ of the density function f based on a function K of a real variable – the so-called kernel of a distribution – and a properly chosen number sequence $\{h_n\}$ from the interval $(0, \infty)$. This estimator of density function of a random variable X under more general assumptions is known in the statistical literature as the Parzen-Rosenblatt estimator or the kernel estimator. The method of kernel estimation presented in the paper has been applied to determine the probability distribution of the groundwater level based on long-term measurements made in the melioration research carried out at the foothill object Długopole.

Key words: probability density function, kernel density estimators, groundwater level

Submitted 29 September 2015, **revised** 15 January 2016, **accepted** 18 April 2016

1. Introduction

In many meteorological and hydrological studies, knowledge of the density function f of the probability distribution of features (random variables) describing the model of the phenomenon allows the determination of the probability of observing the values of these features from a given interval $[a, b]$ according to the formula:

$$Pr\{a \le X \le b\} = \int_a^b f(x)dx$$

There are numerous examples of the many applications of parametric probability distributions for modeling hydrological phenomena (e.g. river flow models, or as in Kuchar et al. 2014) or meteorological phenomena (e.g. models of atmospheric precipitation, cf Kuchar 2004), among which there are: lognormal distribution, two-parameter gamma and beta distributions and generalized gamma distributions (with three parameters). Due to the specific nature of meteorological and hydrological data, the empirical probability distribution determined on the basis of a set of recorded values of given feature X, often shows a large divergence in comparison with known probability distributions. This fact is confirmed by statistical analysis using different statistical tests of hypotheses concerning the verification of the consistency between the empirical distribution of the sample and the hypothetical theoretical distribution. Among the many known methods for estimation of the probability density function f, an often used method is so-called non-parametric kernel estimation.

An important element of this method is the use of a function $K(\cdot)$ of a real variable, called the kernel of a distribution, which satisfies the following condition:

$$\int_{-\infty}^{+\infty} K(x)dx = 1 \tag{1}$$

Next, for a given random sample $X_1, X_2, \ldots, X_n$ we define an estimator $\hat{f}_n$ of the density function f for each $x \in R$, as follows:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h_n^{-1} K\left(\frac{X_i - x}{h_n}\right) \tag{2}$$

where $\{h_n\}$ is a properly chosen number sequence from the interval $(0, \infty)$.

The estimator of density function of a random variable X given by (2) under more general assumptions is called in the statistical literature the Parzen-Rosenblatt estimator or the kernel estimator.

Using the results given by Parzen (1962), Rosenblatt (1956) and by Van Ryzin (1969), we can choose a sequence $\{h_n\}$ and a function K, so that the sequence of estimates $\{\hat{f}_n\}$ is convergent with the probability 1 to the unknown estimated density function f as n tends to infinity. This method of kernel estimation has been applied to determine the probability distribution of the groundwater level based on long-term measurements made in the melioration research carried out at the foothill object Długopole (Gąsiorek et al. 1990).

2. Material and methods

The experiment considered in this article was carried out on the object melioration located in Długopole Zdrój (surface area of about 1.5 hectares) in the district of Kłodzko in the Sudety Mountains, Poland. The district of Kłodzko is characterized by a moderate and mild climate that is favorable for farming and animal breeding, as well as tourism in its various forms. Its characteristic weather is mild winters and slightly cooler summers than in the central part of the Poland. Monthly mean temperatures [in ^{o}C] in the study period (April-September, years 1978-1981) were, consecutively: 5.3, 11.3, 15.3, 14.7, 15.2 and 12.0, while the monthly precipitation totals [in mm] were, respectively: 58.3, 40.3, 88.5, 149.0, 83.8, and 86.6. The melioration study on the foothill object Długopole performed long-term groundwater level measurements using properly installed piezometers (hydrogeological observation holes). Daily registered groundwater levels were averaged, based on measurements from a dozen or so piezometers suitably located at the research station (the experimental data are derived from the Institute of Agricultural and Forest Improvement, cf Gąsiorek et al. 1990). The main objective of this experiment was to determine the probability distribution of the groundwater level within the fixed period of vegetation (from April to September), and then estimation on the basis of the distribution of the average number of days for a given level of ground water. The data set includes the groundwater level measurements listed in rows for specified ranges of levels from 10 to 150 cm every 10 cm (variables p1 to p14) (Table 1). Two experimental years, 1978 and 1979, were very similar in terms of weather conditions and their impact on the size of the runoff and ground water level was also similar. Based on these data, the frequency histogram of the groundwater level was drawn up (Fig. 1). The empirical distribution is the basis for further, more advanced numerical analysis.

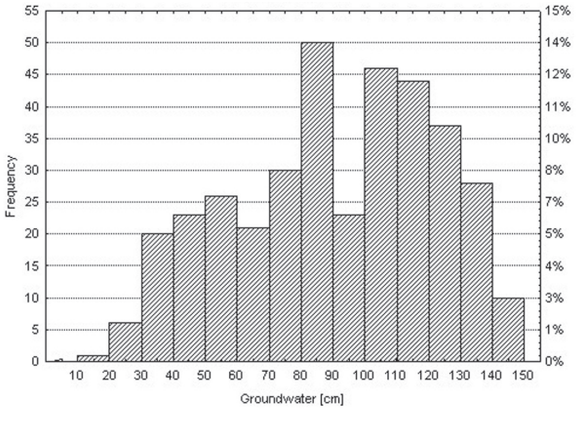

Fig. 1. The empirical probability distribution of groundwater level [cm]

Let us consider an i.i.d. (independent identically distributed) sample X_1, X_2, ..., X_n drawn from a univariate density f, and estimate $\hat{f}_n$ given by (2), where K is the kernel (a real function integrating to one, see (1)), and $h>0$ is the smoothing factor (Akaike 1954; Rosenblatt 1956; Parzen 1962). The fundamental problem in kernel density estimation is that of the joint choice of h and K in the absence of a priori information regarding f. Watson and Leadbetter (1963) show that the choice of h and K should not be split into two independent subproblems. Also, the choice of K largely depends upon the smoothness of f (Devroye 1992). Additionally, we require that the function K is subject to certain conditions of regularity (inter alia, differentiability and integrability). As a result of the condition (1) we have:

$$\int_{-\infty}^{\infty} \hat{f}_n(x)dx = 1$$

and $\hat{f}_n$ satisfies the same conditions of regularities that we superimpose on K. As a kernel K is often assumed density function of the normal distribution $N(0, 1)$ or the density with variance σ^2, i.e. the distribution $N(0, \sigma^2)$. The main problem in issues of kernel estimation of a density function is the optimal choice of the bandwidth h and the kernel K, at which integrated over the mean square error of the kernel estimator would be the lowest for any estimated density, and this means that we are looking for a minimum of risk function as follows:

$$\underset{h,K}{Min}\, R(\hat{f}_n);\ \text{where } R(\hat{f}_n) = E\int_{-\infty}^{\infty}\left[\hat{f}_n(x) - f(x)\right]^2 dx \qquad (3)$$

while h is the width of the window smoothing and K a real function (see e.g. (2)).

Although we know that there are no such values of h and K for which the minimum given by (3) is realized, this analysis of the integrated mean square error for large samples provides useful guidelines in practice. Let us assume that we estimate the density f twice differentiable, for which $\int(f''(x))^2dx<\infty$. Then the asymptotically optimal choice (with $n\to\infty$) of kernel K in a class of symmetric functions and integral in a square is given by the following formula:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right), & -\sqrt{5}\le t\le\sqrt{5} \\ 0 & R/\left[-\sqrt{5};\ \sqrt{5}\right] \end{cases} \qquad (4)$$

It is proved that in the class of the estimated functions, the optimal choice of bandwidth h, asymptotically obtained, expresses the formula (5). From equation (5) it follows that the optimal bandwidth h depends on the kernel K and, unfortunately, on the unknown density f. However,

Table 1. The values of the groundwater level [in cm] with frequency distribution (years: 1978-1979, months: April-September, min = 15 cm, max = 142 cm, total days = 366)

Variable	Interval of variability	Number of days	The values of the groundwater level [cm]
p1	[10; 20]	1	15
p2	(20; 30]	7	28 25 23 26 22 29 29
p3	(30; 40]	20	31 38 37 38 33 39 37 39 35 39 38 34 37 32 34 36 37 31 35 33
p4	(40; 50]	23	42 45 49 43 48 46 46 44 41 47 48 45 49 42 50 41 45 45 49 43 46 46 50
p5	(50; 60]	26	60 52 57 55 53 60 59 54 51 53 57 60 58 58 56 57 57 53 55 51 52 59 51 59 54 53
p6	(60; 70]	21	66 62 68 67 65 63 70 62 65 65 64 69 65 64 66 65 69 64 67 61 62
p7	(70 ; 80]	30	76 74 73 79 74 77 71 74 73 76 72 75 74 74 72 73 77 76 72 74 77 78 79 72 78 76 76 73 76 75
p8	(80; 90]	50	83 82 82 84 90 88 81 90 81 86 85 89 84 86 89 86 84 83 87 89 85 83 82 83 85 84 87 83 82 84 84 89 81 82 87 86 87 88 85 90 88 81 88 87 82 81 85 87 86 88
p9	(90; 100]	23	91 99 97 94 95 96 100 96 93 95 97 92 94 100 98 94 95 100 97 91 95 95 92
p10	(100; 110]	46	101 101 101 101 101 102 102 102 102 102 103 103 103 104 104 105 105 105 105 105 105 105 105 105 105 106 106 106 107 108 109 109 110 110 110 110 110 110 110 104 104 104 106 106 106 110
p11	(110; 120]	44	111 111 111 112 112 112 112 112 112 112 112 113 113 115 115 115 115 115 115 115 115 115 115 115 115 115 114 114 114 116 117 117 118 119 119 120 120 120 120 120 120 120 120 120
p12	(120; 130]	37	121 121 121 121 122 122 122 123 123 123 123 123 123 124 124 124 125 125 125 125 125 125 125 126 127 127 128 128 128 130 130 124 124 124 128 128 128
p13	(130; 140]	28	131 131 132 132 133 133 133 133 133 133 133 134 134 135 135 135 135 135 136 136 137 138 139 139 140 135 135 135
p14	(140; 150]	10	141 141 142 142 142 142 142 142 142 142

$$h = cn^{-1/5},$$
$$\text{where } c = \left\{\int t^2 K(t)dt\right\}^{-2/5} \left\{\int K^2(t)dt\right\}^{1/5} \left\{\int \left[f''(x)\right]^2 dx\right\}^{-1/5} \quad (5)$$

if we consider a sufficiently "rich" family density, e.g. the density of the normal distribution with variance σ^2, and for the kernel K we take the density of the standard normal distribution, then we get:

$$\int \left[f''(x)\right]^2 dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5}, \ \int t^2 K(t)dt = 1, \ \int K^2(t)dt = 0.5\pi^{-1/2} \quad (6)$$

and after consideration of this in (5) we obtain for bandwidth $h = (4/3)^{1/5}\sigma n^{-1/5} \approx 1.06\sigma n^{-1/5}$.

If you use a function K_e given by (4) as the kernel then we obtain $\int t^2 K(t)dt = 1$ and $\int K^2(t)dt \approx 0.27$, and after taking into account these results in (5) for c we obtain that $h \approx 1.05\sigma n^{-1/5}$ (see Gajek, Kałuszka 1994). It is thus seen that the use of the Gaussian kernel leads to practically the same optimum bandwidth. The study of the asymptotic mean square error indicates that the use of a Gaussian kernel instead of the optimal K_e leads to an increase in the error of just a few percent. Therefore, in our further statistical analysis, we used the Gaussian kernel with variance σ^2. Now, let $h_n = \left(\sqrt{n}\right)^{-1}$ and the kernel K be a density function of the normal distribution $N(0, \sigma^2)$, i.e.:

$$K(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-x^2}{2\sigma^2}\right)$$

then the estimator given by (2) has the following form:

$$\hat{f}_n(x) = \frac{1}{\sqrt{2\pi n}\sigma}\sum_{i=1}^{n}\exp\left(\frac{-(X_i - x)^2 n}{2\sigma^2}\right) \tag{7}$$

where $\sigma > 0$ can play the role of the smoothing parameter in finding the optimum width of the smoothing window h, as a function of the number of observations n and parameter σ (e.g. in Gajek, Kałuszka (1994) we have $h \approx 1.06\sigma n^{-1/5}$ or $h \approx 1.05\sigma n^{-1/5}$).

The fundamental argument for using kernel estimators in practice is its unusually important property of strong consistency expressed by the condition of the uniform convergence of estimators $\hat{f}_n$ to an unknown density function f, i.e.:

$$\sup_x \left|\hat{f}_n(x) - f(x)\right| \xrightarrow[n\to\infty]{} 0 \text{ with probability 1} \tag{8}$$

Remarks (cf Wegman 1972):

1.1. The condition (8) holds for a very large class of kernels in real space R if f is uniformly continuous (the result proved by Schuster 1969).

1.2. If both f and a kernel K are continuous, then it is easy to see that (8) defines a random variable for all n.

1.3. If f is a uniformly continuous density and K is a density of bounded variation that satisfies $\lim_{|x|\to\infty}|x|\cdot K(x) = 0$, and if $h_n \xrightarrow[n\to\infty]{} 0$ and $n\cdot h_n^2/\log n \xrightarrow[n\to\infty]{} \infty$ then the condition (8) holds in real space R (the result proved by Naradaya 1965).

For the sake of completeness, we state the following well-known results:

Theorem 2.1 (Naradaya 1965)

If a kernel K is a density on a space $R^m (m \geq 1)$ and f is a uniformly continuous density for the Lebesgue measure, then $\sup_x \left|E\{\hat{f}_n(x)\} - f(x)\right| \xrightarrow[n\to\infty]{} 0$ provided that $h_n \xrightarrow[n\to\infty]{} 0$. (This condition determines the so-called asymptotic unbiasedness).

Theorem 2.2 (Devroye, Wagner 1976)

If a real function K satisfies the following conditions:

(i) K is a probability density on a space $R^m (m \geq 1)$,

(ii) $\sup_x K(x) < \infty$,

(iii) K has compact support, i.e. there exists a $\rho > 0$ such that $\int_{[-\rho,\,\rho]^m} K(x)dx = 1$,

(iv) the closure of the set of discontinuities of K has Lebesgue measure 0,

(v) $h_n \xrightarrow[n\to\infty]{} 0$,

(vi) $h_n^m \cdot n/\log n \xrightarrow[n\to\infty]{} \infty$,

(vii) f is a uniformly continuous density for the Lebesgue measure μ, then $\{\hat{f}_n\}$ is a strongly uniformly consistent for μ.

3. Results

Based on the empirical data of the groundwater level shown in Table 1, and presented in Figure 1 as the frequency histogram, analytically and numerically density function estimates were determined according to the method described in Chapter 2. For different values of the smoothing parameter σ (in the computations, parameter σ ran the interval [0.4; 10] with a step 0.2) the more or less smooth estimated density function $\hat{f}_n$ can be obtained (for selected values of the parameter σ we show: for $\sigma = 5$ see Fig. 2, for $\sigma = 7$ see Fig. 3 and for $\sigma = 10$ see Fig. 4).

An additional criterion for assessing the effectiveness (goodness of fit) of estimators can be to maximize the probability Pr as follows:

$$\max_{\sigma>0} Pr\{\chi^2_{k-1} > \chi^2_{obl}\}, \text{ where } \chi^2_{obl} = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \tag{9}$$

O_i and E_i are the observed and the expected values for the i-th class of analyzed variable ($i = 1,\ldots, k$), respectively and χ^2_k is a random variable that has a central chi-squared distribution with k degrees of freedom. It is worth noting the expected values $E_i = n \cdot p_i$, where n is the total number of observations (in our case $n = 366$ days) and p_i expresses the probability for the i-th variation interval for $i = 1,\ldots, k$. (in our case $k = 14$). For the chosen values of smoothing parameter σ we received appropriate p-values for the test statistic χ^2 given by formula (9), and so, for $\sigma = 5$, 7 and 10 we have $Pr\{\chi^2_{13} > 5.89\} \approx 0.95$, $Pr\{\chi^2_{13} > 6.32\} \approx 0.94$ and $Pr\{\chi^2_{13} > 6.18\} \approx 0.934$, respectively.

4. Discussion

The results obtained confirm the high usefulness of the method of kernel estimation to determine the probability distribution of the groundwater level and to estimate the most probable number of days for a given level of ground-

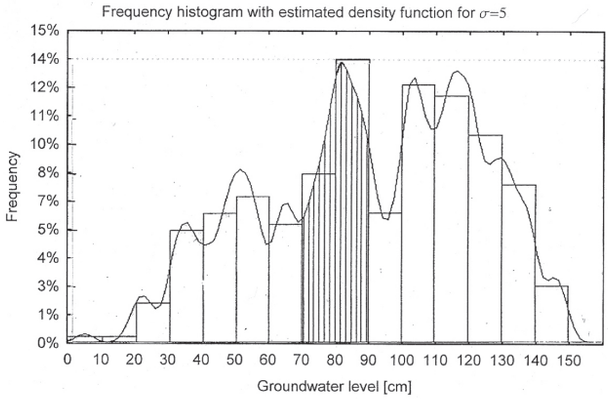

Fig. 2. Frequency histogram for the source data of groundwater with estimated density function (the smoothing parameter $\sigma = 5$)

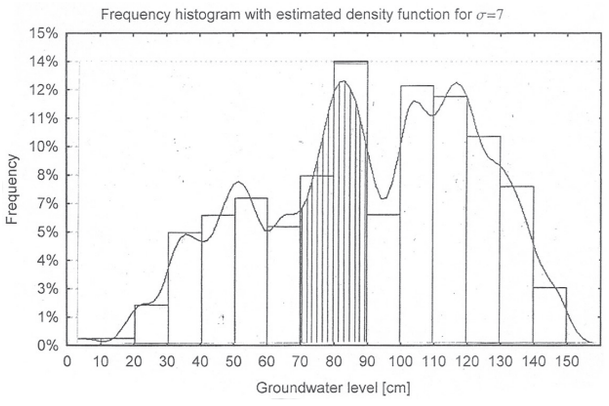

Fig. 3. Frequency histogram for the source data of groundwater with estimated density function (the smoothing parameter $\sigma = 7$)

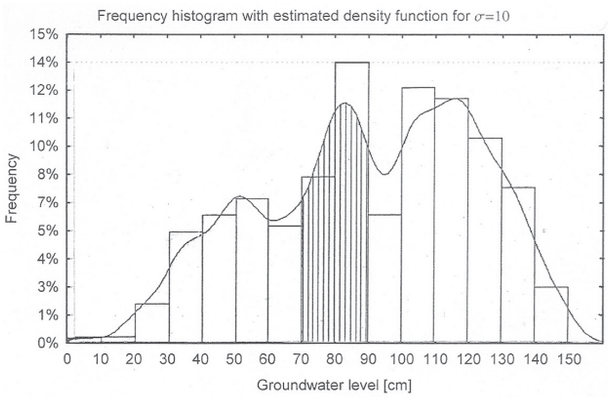

Fig. 4. Frequency histogram for the source data of groundwater with estimated density function (the smoothing parameter $\sigma = 10$)

water in the studied vegetation period at a melioration object (in our case at the foothill object Długopole). For example, for values of the groundwater level in the range 70-90 cm, on the basis of long-term measurements, the predicted (expected) number of days is 77 for the estimated density function with smoothing parameters $\sigma = 5$ and $\sigma = 7 (Pr\{x \in [70, 90]\} \approx 0.21)$, and 73 days when we use the estimator $\hat{f}_n$ with smoothing parameter $\sigma = 10 (Pr\{x \in [70, 90]\} \approx 0.20)$, while the observed number of days for this interval was 80. As an alternative for choosing the smoothing factor h in the Akaike-Parzen-Rosenblatt density estimate, Devroye (1989) introduced the double kernel estimate and its usefulness was demonstrated in extensive simulation studies in Berlinet and Devroye (1994). Among the publications on the problems of estimating the unknown density function, one should consider monograph by Silverman (1986), in which the author gives several proposals for choosing the window width. Another important but much more difficult approach to the problem of stochastic modeling of hydrological or meteorological data is to use methods of multivariate density function estimation (see Scott 1992). In our problem, we can consider simultaneously with the level of groundwater both the quantity of precipitation and mean daily temperatures.

Acknowledgments

The author is grateful to the reviewers for their very helpful comments and suggestions.

Bibliography

Akaike H., 1954, An approximation to the density function, Annals of the Institute of Statistical Mathematics, 6 (2), 127-132, DOI: 10.1007/BF02900741

Berlinet A., Devroye L., 1994, A comparison of kernel density estimates, Publications de l'Institut de Statistique de l'Université de Paris, XXXVIII – Fascicule, 3, 3-59

Devroye L., 1989, The double kernel method in density estimation, Annales de l'Institut Henri Poincaré, 25, 533-580

Devroye L., 1992, A note on the usefulness of superkernels in density estimation, Annals of Statistics, 20 (4), 2037-2056

Devroye L., Wagner, T.J., 1976, Nonparametric discrimination and density estimation, Technical Report 183, Electronic Research Center the University of Texas at Austin, TX, USA

Gajek L., Kałuszka M., 1994, Statistical Inference, Wydawnictwo Naukowo-Techniczne, Warsaw, 304 pp., (in Polish)

Gąsiorek E., Michalski A., Pływaczyk A., 1990, Analysis of land improvement objects data, Zeszyty Naukowe Akademii Rolniczej we Wrocławiu, 65-76, (in Polish)

Kuchar L., 2004, Using WGENK to generate synthetic daily weather data for modeling of agricultural processes, Mathematics and Computers in Simulation, 65 (1-2), 69-75, DOI: 10.1016/j.matcom.2003.09.009

Kuchar L., Iwański S., Jelonek L., Szalińska W., 2014, Application of spatial weather generator for the assessment of climate change impacts on a river runoff, Geografie, 119 (1), 1-25

Nadaraya E.A., 1965, On nonparametric estimation of density functions and regression curves, Theory of Probability and Its Applications, 10 (1), 186-190, DOI: 10.1137/1110024

Parzen E., 1962, On the estimation of a probability density function and the mode, Annals of Mathematical Statistics, 33 (3), 1065-1076, DOI: 10.1214/aoms/1177704472

Rosenblatt M., 1956, Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics, 27 (3), 832-837, DOI:10.1214/aoms/1177728190

Schuster E.F., 1969, Estimation of a probability density function and its derivatives, The Annals of Mathematical Statistics, 40 (4), 1187-1195

Scott D.W., 1992, Multivariate density estimation. Theory, Practice and Visualization, John Wiley & Sons Inc., New York, USA, 317 pp., DOI: 10.1002/9780470316849

Silverman B.W., 1986, Density estimation for statistics and data analysis, CRC Press, London, UK, 176 pp.

Van Ryzin J., 1969, On strong consistency of density estimates, Annals of Mathematical Statistics, 40 (5), 1765-1772

Watson G.S., Leadbetter M.R., 1963, On the estimation of the probability density, Annals of Mathematical Statistics, 34 (2), 480-491

Wegman E.J., 1972, Nonparametric probability density estimation, I: A summary of a variable methods, Technometrics, 14 (3), 533-546, DOI: 10.1080/00401706.1972.10488943